SAN FRANCISCO STATE UNIVERSITY

A Visual Analytics Tool to Visualize Data Summary View using Sampled Data

by

Poornank Purohit

Supervisors:

Dr. Shahrukh Humayoun

Dr. Jingyi Wang

A thesis submitted in partial fulfillment for the
degree of Master of Science in Computer Science

in the
Computer Science
Department of Computer Science

December 2023

Declaration of Authorship

I, Poornank Purohit, declare that this project report titled, 'A Visual Analytics Tool to Visualize Data Summary View using Sampled Data' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research project at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution."

— Albert Einstein

SAN FRANCISCO STATE UNIVERSITY

Abstract

Computer Science

Master of Science in Computer Science

by Poornank Purohit

This applied research project delves into advanced data visualization techniques using D3.js, React, and JavaScript, aimed at analyzing large datasets and providing summary views. The focus is on crafting dynamic, interactive visualizations like scatter plots, line charts, and histograms to efficiently depict complex data. We address challenges such as integrating D3.js with React's virtual DOM and enhancing performance for substantial datasets. A significant aspect of our research is the implementation of various sampling methods, including random sampling, stratified sampling, and systematic sampling. These techniques are pivotal in ensuring accurate and insightful data representation, with random sampling providing an unbiased overview, stratified sampling enabling targeted analysis with increased precision, and systematic sampling proving effective in visualizing ordered data like time-series. The inclusion of these sampling methods enhances the efficacy and efficiency of our visualizations, making them more representative and insightful. This study contributes to the field by demonstrating practical applications of these technologies in data visualization and offers insights for future research and methodological advancements. The project holds substantial value in improving the comprehension and analysis of big data and lays a groundwork for similar future endeavors in the field of data visualization.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Shahrukh Humayoun and Dr. Jingyi Wang, their continuous support and guidance was very helpful in all times while working on this research project and, I got to learn how to use different tools and technologies to make this project possible.

I would also like to extend my heartfelt appreciation to my family for their unwavering support, encouragement, and love. Their belief in me has been the driving force behind my success, and I cannot thank them enough.

Finally, I would like to thank my peers for their continuous support and valuable input throughout this project. Their constructive feedback, encouragement, and ideas have been invaluable.

Contents

List of Figures

Dedicated to my beloved Parents and Brother

Chapter 1

Introduction

1.1 The Dynamic Landscape of Data Visualization

Data visualization, especially in the context of summarizing large datasets, has seen a remarkable evolution, becoming crucial for interpreting and communicating complex data. The role of data visualization in simplifying and elucidating large volumes of information is highlighted in Healy and Moody's "Data Visualization in Sociology," emphasizing its central role in making complex datasets comprehensible and actionable [1]. This evolution is driven by the need to distill meaningful insights from large datasets, transforming decision-making processes across various sectors.

Edward Tufte's seminal work, "The Visual Display of Quantitative Information," underscores the transformative power of visual representation, stressing the need for clarity and precision in the depiction of data [2]. This exploration serves to contextualize the significance of this research within the wider evolution of data visualization, emphasizing the growing importance of summary views in understanding complex data.

The exponential growth in demand for innovative tools and platforms in data visualization, particularly for summarizing large datasets, guides the direction of this research project. This study responds to these evolving challenges by aiming to develop a robust and adaptable approach to visually interpreting complex data, evolving in line with the expanding dimensions of information.

Anderson's "Introduction to Random Sampling" informs the project's motivation, highlighting the limitations inherent in traditional data visualization tools when faced with the need for dynamic exploration of large datasets [3]. This research dissects these limitations, aiming to methodologically and theoretically contribute to the broader field of data visualization,

particularly in the creation of summary views. The overarching goal of this research is to advance the field of data visualization, with a specific focus on summarizing large datasets.

1.2 Evolution of Techniques for Summary View of Large Datasets

The evolution of techniques for summarizing large datasets has been a cornerstone of effective data analysis. Cochran's "Sampling Techniques" illustrates the fundamental role these methods have played in statistical analysis and data interpretation [4]. This project examines the transition from traditional techniques to more advanced, adaptive strategies for summarizing large datasets, highlighting their importance in contemporary data-driven environments.

Technological advancements have revolutionized methods for summarizing large datasets. In big data contexts, where the volume and variety of data challenge conventional techniques, Mahmud et al. emphasize the need for evolving methods to keep pace with technological advancements [5]. This project responds to these challenges by demonstrating how advanced techniques can enhance the capacity to extract meaningful insights from large volumes of data.

In an era dominated by big data, efficient techniques for summarizing large datasets are crucial. Insights from Healy and Moody on the evolution of data visualization inform this project's approach to integrating new techniques for creating summary views [1]. The project aims to demonstrate the effectiveness of these methods in enhancing data visualization and interpretation.

Keim et al.'s work on "Visual Analytics: Definition, Process, and Challenges" provides a framework for understanding the intricate relationship between data analysis techniques and visual representation, particularly in the context of summarizing large datasets [6]. This project explores various advanced strategies like systematic and stratified sampling, assessing their effectiveness in different data contexts.

The project delves into the interplay between different techniques for summarizing large datasets and data interpretation. Lohr's "Sampling: Design and Analysis" serves as a guide, emphasizing the critical role of these techniques in interpreting and making decisions based on large datasets [7].

In summary, Section 1.2 sets the stage for understanding the evolution and modern application of techniques for summarizing large datasets. It underscores the project's commitment

to advancing data visualization through the integration of sophisticated strategies, addressing the challenges of big data, and enhancing the interpretability and utility of data-driven insights.

1.3 Objectives and Scope of the Research in Summary View of Large Datasets

The principal objective of this research is to examine and enhance the integration of advanced sampling techniques with data visualization techniques for creating effective summary views of large datasets. This involves a detailed examination of the efficiency of various sampling methods, their impact on data interpretation, and the development of visualization tools that can represent complex data concisely and effectively. The project aims to bridge the theoretical aspects of statistical methods with practical data visualization applications, highlighting the significance of this integration in the context of large datasets [1].

The scope of this research encompasses an in-depth analysis of sampling techniques ranging from traditional methods like simple random sampling to more intricate strategies such as stratified and systematic sampling [4]. These concepts are then extended to the practical realm of data visualization, where these methods are applied to create summary views of large datasets. The integration of these sampling methods into visualization tools, reflecting the principles of effective data presentation, is a key focus [8].

Given the complexities and volume of big data, implementing efficient sampling techniques for data visualization is critical. This research addresses these challenges by leveraging the insights and advancements in data visualization and sampling methods [5]. The goal is to develop visualization tools that provide not only comprehensive analyses but are also intuitive and user-friendly, catering to the needs of diverse users.

The methodological approach of this research is grounded in a mix of theoretical study and practical application. It includes a thorough review of existing literature on sampling techniques and data visualization, followed by the development and testing of visualization tools suited for summarizing large datasets. This approach ensures academic rigor and practical relevance [9].

The project places a strong emphasis on innovation in data visualization. By incorporating advanced sampling techniques with contemporary visualization tools, the research aspires to push the boundaries in the field of data visualization, particularly in summarizing large datasets [2].

The research is poised to make significant contributions to the field of data visualization and analytics. The tools and methodologies developed are intended to be versatile, applicable in various contexts beyond the specific scope of this project, potentially influencing the broader trajectory of the field [6].

1.4 Research Methodology and Framework for Summary View Visualization

The methodological framework of this research is designed to rigorously explore and implement advanced sampling techniques in the context of data visualization for summary views of large datasets. This approach combines theoretical study with practical experimentation to create a bridge between statistical concepts and real-world application.

The research begins with an extensive literature review, focusing on seminal works in sampling methods and data visualization. This phase involves a deep dive into Cochran's "Sampling Techniques" [4] and Tufte's "The Visual Display of Quantitative Information" [2], laying a solid theoretical foundation for the project. Contemporary studies like those by Healy and Moody are also reviewed to align the research with the latest developments [1].

Following the theoretical groundwork, the focus shifts to developing various sampling techniques. This phase adapts traditional methods and explores advanced strategies suitable for summarizing large datasets. The development process adheres to principles from established literature, tailored to the specific challenges of data visualization [10].

The core of the methodology involves the practical implementation of these sampling techniques in data visualization tools. These tools are developed with a focus on handling large datasets, offering intuitive interfaces for data exploration and analysis, as advocated by Murray [8].

The project aligns the developed tools with contemporary trends in data visualization. Current research, such as the work of Heer and Shneiderman, informs the design and functionality of the tools, ensuring they are statistically sound and align with modern visualization practices [11].

The research adopts an interdisciplinary approach, combining statistics, computer science, and domain-specific knowledge. This approach ensures a comprehensive understanding of the technical and contextual aspects of data visualization for summarizing large datasets.

1.5 Significance of the Study

This research holds substantial relevance in the field of data visualization, particularly in the context of summarizing large datasets. The integration of advanced sampling techniques with interactive visualization tools addresses a vital need for efficient and effective representation of complex data. This contribution is particularly pivotal given the increasing reliance on data-driven decision-making in various sectors, aligning with Healy's perspectives on practical data visualization [9].

The exploration and application of diverse sampling methods, such as stratified random and systematic sampling, mark a significant advancement in data analysis methodologies. This aspect of the study extends beyond academic significance, enhancing the accuracy and interpretability of data visualizations, thereby making complex datasets more accessible and actionable [4].

The study has the potential to transform the way large datasets are analyzed and visualized. By developing tailored visualization tools based on sophisticated sampling strategies, it enables a deeper and more nuanced understanding of large volumes of data. This aligns with Park, Cafarella, and Mozafari's work on visualization-aware sampling for large databases [12].

Academically, the study adds to the growing research at the intersection of data science, statistics, and visualization. It provides empirical evidence and methodological insights that can be leveraged in future research, particularly in areas requiring analysis of large-scale data, as discussed in "Big Data: A Survey" by Chen, Mao, and Liu [13].

Practically, the research enhances the capabilities of existing analytical tools. By offering a more efficient and effective way to analyze and visualize large datasets, the project supports informed decision-making, in line with the trend towards data-driven approaches across various industries.

The methodologies and tools developed in this study have implications beyond specific data visualization contexts. The principles and techniques can be adapted and applied across different domains, potentially leading to innovations in data visualization and analysis in various industries.

1.6 Project Structure

This project is structured to provide a comprehensive exploration of advanced sampling techniques integrated with data visualization for summarizing large datasets. Each chapter

builds upon the previous, creating a cohesive narrative that effectively communicates the research findings.

Chapter 1: Introduction This chapter introduces the background, objectives, and significance of the study. It provides a contextual understanding of the evolution of data visualization and the significance of summary views in interpreting large datasets.

Chapter 2: Sampling Techniques in Data Analysis This chapter provides a detailed examination of various sampling techniques crucial for summarizing large datasets. It aims to offer a thorough understanding of each technique, its application, and its relevance to data visualization.

Chapter 3: Review of Previous Work and Methodological Integration This chapter reviews previous work in the area of summary views of large datasets, drawing on research papers and studies. It discusses the integration of these methods into the current project, illustrating the novel approach taken to address challenges in data visualization.

Chapter 4: Visualization Tools and Techniques This chapter focuses on the visualization tools and techniques utilized in the project, detailing their development and implementation for enhanced data analysis and interpretation.

Chapter 5: Conclusions and Future Work This chapter presents the study's findings, including a detailed analysis of the data visualizations produced. It assesses their effectiveness and practical applications, concluding with future research directions and the study's overall contribution to data visualization.

Chapter 2

Sampling Techniques in Data Analysis and Visualization

2.1 Overview of Sampling Techniques

Sampling techniques are fundamental to the field of data analysis, providing a means to draw meaningful conclusions from large datasets. The essence of sampling lies in selecting a representative subset of a population, thereby enabling the analysis of data attributes without the need to examine the entire dataset. This section provides an overview of various sampling techniques and their significance in data analysis.

The evolution of sampling techniques has been closely tied to the development of statistical theory and practice. As Cochran illustrates in "Sampling Techniques," the history of sampling methods is rich and varied, reflecting the growing complexity of datasets and the need for more refined analytical approaches [14].

Simple random sampling is one of the most basic yet powerful methods. It involves selecting a subset of a population in which each member has an equal probability of being chosen. This technique is fundamental to ensuring unbiased and representative samples, as noted by Lohr in "Sampling: Design and Analysis" [6].

Stratified sampling is a method where the population is divided into homogenous subgroups, or strata, and samples are drawn from each group. This approach ensures that specific characteristics of the population are represented in the sample, as discussed in Sedgwick's "Stratified Cluster Sampling" [15].

Systematic sampling involves selecting data points at regular intervals from an ordered list. It combines elements of randomness and uniformity, making it particularly useful for datasets with an inherent order, as detailed in Anderson's "Introduction to Random Sampling" [5].

In cluster sampling, the population is divided into clusters, and a sample of these clusters is selected for analysis. This method is effective when it is impractical or impossible to study the entire population, offering a practical alternative as illustrated by Mahmud et al. in their work on data partitioning and sampling methods [16].

While sampling techniques provide a viable solution for analyzing large datasets, they come with challenges. Selecting a truly representative sample, dealing with biases, and determining the appropriate sample size are critical considerations. Addressing these challenges requires careful planning and a deep understanding of statistical principles, as emphasized in the works of Cochran and Lohr [14] [6].

Sampling techniques are integral to data analysis, offering a means to gain insights from large datasets in a practical and efficient manner. The choice of sampling method depends on the nature of the dataset, the objectives of the analysis, and the specific characteristics of the population being studied. As data continues to grow in size and complexity, the role of sampling in data analysis becomes increasingly important, driving the need for innovative and adaptive sampling strategies.

2.2 Visualization in Data Analysis

Visualization plays a crucial role in data analysis by transforming complex data sets into comprehensible visual formats. This section introduces the concept of data visualization, focusing on its interplay with sampling techniques to provide summary views of large datasets.

2.2.1 Traditional Data Visualization

Traditional data visualization encompasses a range of techniques and tools designed to represent data in a visual context, such as charts, graphs, and maps. The primary goal is to communicate information clearly and efficiently, making complex data accessible to a wider audience. As Tufte discusses in "The Visual Display of Quantitative Information," effective visualization hinges on presenting data in a way that emphasizes true meaning without distorting what the data has to say [1].

In traditional visualization tools like bar graphs, pie charts, and line plots, sampling techniques play a vital role in summarizing and presenting large datasets. Simple random

sampling, for example, is often used to select a representative subset of data for visualization, ensuring that the resulting graphics are both manageable in size and reflective of the larger dataset.

Various software and tools have been developed for traditional data visualization. Tools like Microsoft Excel, Tableau, and IBM SPSS offer functionalities for creating static visualizations. These tools often incorporate basic sampling methods to manage large datasets, allowing users to generate visual summaries that capture key data trends. As Heer and Shneiderman note in "Interactive Dynamics for Visual Analysis," these tools support the fluent and flexible use of visualizations, albeit within the constraints of static, predefined formats [2].

While traditional visualization tools are effective for certain types of data, they may fall short when dealing with extremely large or complex datasets. Limitations in scalability and interactivity can hinder the ability to extract meaningful insights, particularly when dynamic, real-time data exploration is required. These limitations underscore the need for more advanced visualization techniques and tools, capable of handling larger volumes of data with greater flexibility.

Traditional data visualization tools have set the foundation for presenting data visually, but the growing complexity and size of datasets necessitate more advanced approaches. The integration of sampling techniques in these tools highlights their importance in creating effective visual summaries, but also points to the limitations that emerging trends in visualization seek to address.

While these tools have undeniably played a pivotal role, a critical examination of their limitations becomes imperative. The section aims to balance the accolades with a transparent discussion on the challenges these traditional tools pose, particularly concerning the integration of advanced sampling and sorting techniques. Infographics can serve as powerful tools to visually compare the customization options offered by these tools against the envisioned capabilities of the proposed novel application. By highlighting the identified gaps and opportunities for innovation, this section aims to set the stage for the subsequent chapters that delve into addressing these limitations.

2.2.2 Advancements in Visualization Techniques

The landscape of data visualization is continuously evolving, with emerging trends focusing on enhanced interactivity, scalability, and the integration of advanced analytics. As datasets grow in size and complexity, there is a growing need for visualization techniques that can not only summarize large datasets effectively but also provide dynamic, real-time insights.

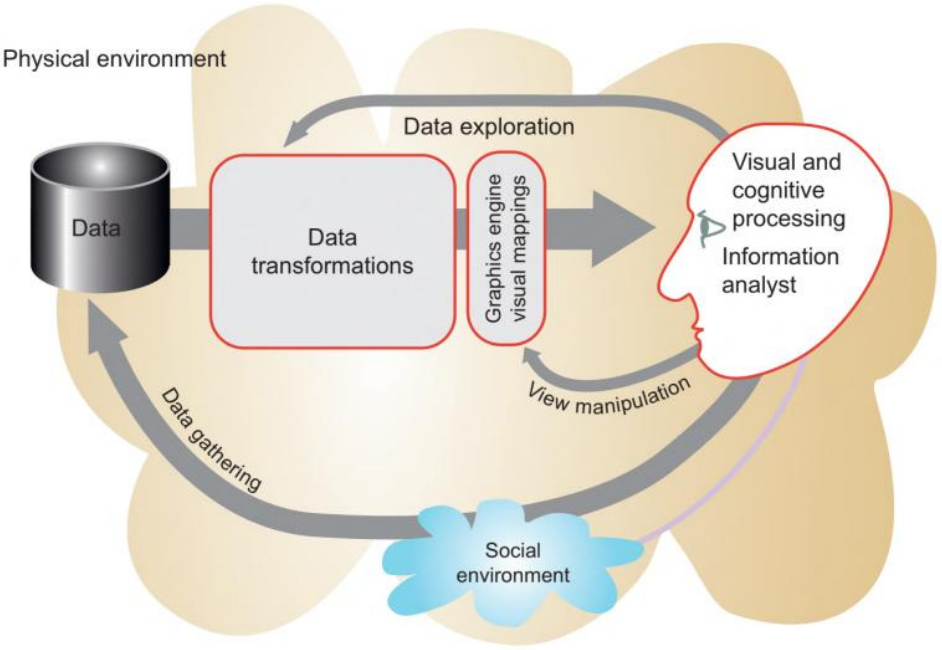

FIGURE 2.1: Visualization facilitates hypothesis formation. For example, the visualization led to questions about how the pockmarks might have formed and motivated a research paper concerning the geological significance of the features (Gray, Mayer, Hughes Clarke, 1997)
[14]

Modern data visualization trends incorporate advanced sampling techniques to manage and interpret large datasets efficiently. Techniques like stratified and cluster sampling are increasingly important in visualizations dealing with big data, as they allow for the representation of diverse data characteristics while managing the data volume effectively. As noted by Park, Cafarella, and Mozafari, visualization-aware sampling for large databases is becoming essential in the era of big data [8].

New visualization tools and platforms are emerging, leveraging the power of machine learning and artificial intelligence to provide deeper insights into data. Tools like D3.js and Google's Data Studio represent this trend, offering more flexibility and customization in data representation. These tools often integrate sophisticated sampling methods to create interactive and dynamic visualizations that can handle large volumes of data efficiently, as discussed by Murray in "Interactive Data Visualization for the Web" [12].

One of the key trends in modern data visualization is the focus on user experience and interactivity. Interactive dashboards, real-time data streams, and user-driven data exploration are becoming more prevalent. This shift is aligned with the principles outlined by Heer and Shneiderman, where interactive dynamics for visual analysis are emphasized to support a more fluent and flexible use of visualizations [2].

With the advent of big data and the Internet of Things (IoT), visualization tools are increasingly required to process and display large volumes of data from various sources. The challenge lies in not only representing this data effectively but also in ensuring that it is accessible and actionable for users. The integration of advanced sampling techniques in these tools is crucial for summarizing vast datasets and providing meaningful visual insights.

Emerging trends in data visualization are marked by a shift towards more interactive, scalable, and analytics-driven approaches. The integration of advanced sampling techniques in these new visualization tools is critical for effectively summarizing and interpreting large datasets. As visualization technologies continue to evolve, they offer promising avenues for more insightful and dynamic data exploration.

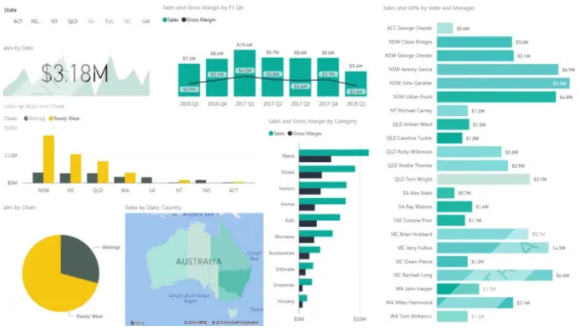

FIGURE 2.2: Analysis and Visualization of data created in Tableau

The figure 2.2 illustrates the analysis and visualization of data created in Tableau, showcasing the advanced capabilities of this tool in simplifying complex datasets. By leveraging Tableau's user-friendly interface and robust features, analysts can craft visually compelling dashboards, revealing hidden insights and trends in the data. This visualization serves as an exemplar of how modern interactive platforms, informed by the works of Healy and Moody, have evolved from early graphing techniques to enable more insightful hypothesis formation and data exploration [1].

The chapter further explores how these emerging trends align with the overarching goals of advancing the field. While the emphasis on visualization grammar is evident, the section probes into the implications of this focus on user experience and interactivity. Plotly, for instance, places a premium on creating web-based interactive visualizations, raising questions about the potential impact on user engagement and understanding. By dissecting these considerations, the section seeks to engage the reader in a thoughtful exploration of the evolving dynamics in data visualization methodologies.

2.3 Sampling Techniques in Data Analysis

2.3.1 Introduction

In the realm of data visualization and analysis, especially in complex fields like visual analytics, the role of sampling techniques cannot be overstated. This project, focused on data visualization, relies heavily on robust sampling methods to effectively manage and interpret large datasets. Sampling, in this context, is not just a statistical necessity but a strategic tool to glean meaningful insights from vast amounts of data.

The development and application of the JavaScript-based tools in this project, such as 'Histogram.js' and 'ScatterPlot.js', necessitate a deep understanding of various sampling techniques. These tools are designed to not only display data but also to provide a mechanism for users to interact with and understand the complexities of analytical statistics through various sampling methods. As we delve deeper into the specific sampling techniques used in the project, it is important to recognize their significance in enhancing the efficacy and efficiency of data visualization.

The subsequent sections will explore different sampling techniques utilized in the project, starting with an overview of random sampling, followed by an in-depth look at specific methods such as simple random sampling, cluster random sampling, stratified sampling, stratified random sampling, and systematic sampling. Each of these techniques contributes uniquely to the handling and representation of analytical data, ensuring that the visualizations are both accurate and insightful.

This introduction sets the stage for a detailed exploration of how sampling techniques are intricately woven into the fabric of the project's data visualization tools, thereby enriching the overall analysis and understanding of analytical data.

2.3.2 Random Sampling Technique

Random sampling is a fundamental statistical technique used across various research fields for data analysis and visualization. This method involves selecting a subset from a larger population, where each member has an equal probability of being chosen. Anderson emphasizes the significance of random sampling in providing unbiased and representative results in "Introduction to Random Sampling" [5].

The algorithm for random sampling employed in this project can be outlined in the following steps:

1. Defining the Population: Identify the entire set of analytical data from which the sample will be drawn. 2. Specifying the Sample Size: Determine the number of data points to be included in the sample. 3. Random Selection: Use a random mechanism (like a random number generator) to select data points from the population. Each member of the population should have an equal probability of being selected. 4. Creating the Sample: Compile the randomly selected data points to form the sample.

The process of random sampling can be algorithmically represented as follows:

Algorithm 1 Random Sampling Technique

```
procedure RandomSampling(data, sampleSize)
    n ← length(data)
    Initialize sampledData as an empty list
    for i ← 1 to sampleSize do
        index ← RandomInteger(1, n)
        Append data[index] to sampledData
    end for
    return sampledData
end procedure
```

Random sampling is crucial in data visualization, especially when dealing with large datasets. It allows for the creation of summary views that accurately reflect the larger dataset without the need for exhaustive data rendering. This technique is particularly beneficial in big data visualization, as noted in "Big Data: A Survey" by Chen, Mao, and Liu [17], and in the work of Park, Cafarella, and Mozafari on visualization-aware sampling for large databases [8].

Cochran's "Sampling Techniques" provides a comprehensive foundation on the statistical principles underpinning random sampling, highlighting its role in ensuring representativeness and reducing biases [14]. The random sampling method is valued for its straightforwardness and theoretical simplicity, making it a preferred choice in many preliminary data analyses. In practice, random sampling is implemented through various means, such as using random number generators or other algorithmic approaches. This ensures that every element of the population has an equal chance of being selected, as detailed by Lohr in "Sampling: Design and Analysis" [6].

The main advantage of random sampling lies in its ability to produce unbiased samples, crucial for valid statistical inferences. However, it can be challenging to implement effectively in very large populations and may require significant computational resources for large datasets.

2.3.3 Simple Random Sampling

Simple Random Sampling is a fundamental statistical method widely used across various research disciplines. This technique involves selecting a sample from a larger population in a way that each member has an equal chance of being included. It is the most basic form of random sampling, ensuring unbiased representation of the population.

Cochran's "Sampling Techniques" provides a detailed exploration of simple random sampling, highlighting its importance in achieving statistical representativeness [14]. The simplicity of this method makes it a fundamental tool in statistical analysis, as it avoids the biases inherent in non-random sampling techniques.

Simple Random Sampling can be algorithmically represented as follows:

Algorithm 2 Simple Random Sampling Technique

```
procedure SimpleRandomSampling(data, sampleSize)
    n ← length(data)
    Initialize sampledData as an empty list
    for i ← 1 to sampleSize do
        index ← RandomInteger(1, n)
        Append data[index] to sampledData
    end for
    return sampledData
end procedure
```

Simple random sampling has wide applications in data analysis and visualization. Its use is crucial in fields such as market research, clinical trials, and environmental studies, where unbiased representation is essential. In data visualization, particularly in generating summary views of large datasets, this technique ensures that the visual representation accurately reflects the overall characteristics of the entire dataset.

In the realm of big data, as discussed in "Big Data: A Survey" by Chen, Mao, and Liu [17], simple random sampling is instrumental in managing the volume and complexity of data. It allows analysts to extract manageable subsets from large datasets for visualization and analysis, thereby making the process more efficient and computationally feasible.

Lohr's "Sampling: Design and Analysis" emphasizes the statistical rigor provided by simple random sampling in ensuring the reliability of data analysis [6]. This technique forms the basis of many statistical inference methods, serving as a cornerstone for hypothesis testing and data interpretation.

While simple random sampling is straightforward, it can be challenging to implement effectively in extremely large or inaccessible populations. Additionally, this technique requires

careful consideration of sample size to ensure that the results are statistically significant and representative of the population.

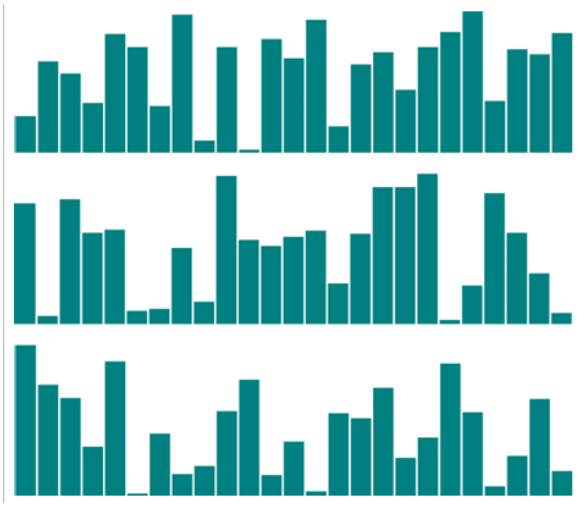

FIGURE 2.3: Bar Chart with random values
[18]

In this figure, a bar chart with random values is presented, highlighting the implementation of Simple Random Sampling in the Visual Analytics Tool. Following Neyman's principles, it ensures unbiased and representative visualizations by giving each data element an equal chance of selection. This visual representation underscores the tool's commitment to statistical rigor and effective data visualization [19].

The evolution of simple random sampling has been marked by its integration with modern data analysis tools and technologies. Contemporary data visualization platforms often incorporate this sampling method to provide dynamic and interactive visual representations of data, enhancing user engagement and understanding.

Simple random sampling remains a critical technique in the toolkit of researchers and analysts. Its ability to provide unbiased and representative samples makes it invaluable in a wide range of applications, from traditional statistical analysis to modern data visualization and big data analytics.

2.3.4 Cluster Random Sampling

Cluster random sampling is a method used in statistical analysis to sample a population when individual sampling is impractical due to size or geographical distribution. It involves dividing the population into clusters, then randomly selecting whole clusters for analysis. This technique is particularly useful for large-scale surveys and geographical studies.

Cluster sampling's theoretical underpinnings are discussed in detail in Cochran's "Sampling Techniques" [14]. This method is essential for handling large populations spread over wide areas, where simple random sampling might be too costly or logistically challenging.

Algorithmic Representation: The process of cluster random sampling can be algorithmically represented as follows:

Algorithm 3 Cluster Random Sampling Technique

```
procedure ClusterRandomSampling(data, numClusters)
    Divide data into N clusters
    Randomly select numClusters from N
    Initialize sampledData as an empty list
    for each selected cluster do
        Append all data from the cluster to sampledData
    end for
    return sampledData
end procedure
```

Cluster random sampling is widely used in fields like epidemiology, market research, and environmental studies. In epidemiology, it helps in studying disease prevalence in different regions by selecting clusters like towns or neighborhoods. Market researchers use it to understand consumer behavior across different geographical clusters.

In the context of big data, as highlighted in "Big Data: A Survey" by Chen, Mao, and Liu [17], cluster random sampling is valuable for managing and analyzing large datasets, especially when data points are naturally grouped. In data visualization, this method allows for the creation of summary views representing different clusters, aiding in the identification of patterns and trends across various segments of the data.

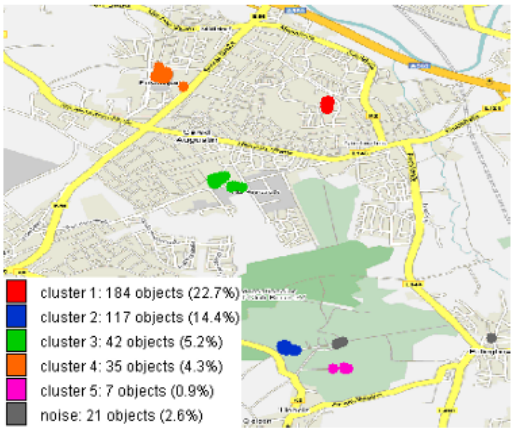

Figure 2.4: Positions of stops have been extracted from the database. By means of clustering, frequently visited places have been detected. [9]

This figure illustrates the application of Cluster Random Sampling in the Visual Analytics Tool, showcasing the positions of stops extracted from a database. By clustering frequently visited places, it highlights the efficiency of this sampling technique in dealing with geographically dispersed or clustered datasets. The figure underscores the strategic approach of Cluster Random Sampling in visualizing complex data structures [20].

In the context of analytics, cluster random sampling allows for the efficient handling of extensive datasets that are naturally segmented, such as data categorized by leagues, age groups, or geographical regions. By analyzing representative clusters, the project can glean insights that are reflective of the broader population, while minimizing the computational load.

Cluster random sampling is built on the principle that certain population characteristics can be captured by fully analyzing subsets (clusters) of that population. This approach is informed by the works of renowned statisticians and their studies on sampling techniques, providing a strong theoretical foundation for its application in data analysis.

One of the main challenges of cluster random sampling is the potential for increased sampling error compared to simple random sampling. This can occur if clusters are not homogenous. Sedgwick's work on "Stratified Cluster Sampling" provides insights into addressing these challenges [15].

Lohr's "Sampling: Design and Analysis" emphasizes the importance of carefully selecting clusters to ensure that the sample is representative of the population [6]. The reliability of conclusions drawn from cluster random sampling depends on the homogeneity and the random selection of clusters.

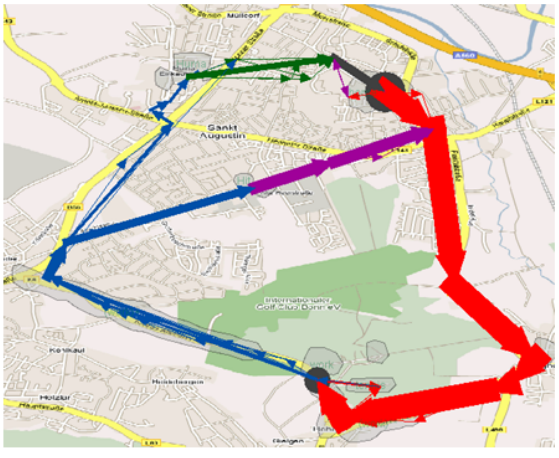

FIGURE 2.5: A result of clustering and summarization of movement data: the routes between the significant places
[9]

The evolution of cluster random sampling has been marked by its adaptation to modern research needs, including its application in digital and social media analytics. As datasets become increasingly complex, cluster sampling offers an efficient way to extract meaningful information without analyzing the entire dataset.

Cluster random sampling remains a vital technique in statistical analysis, particularly for large and geographically dispersed populations. Its application spans various domains, proving essential in drawing accurate and reliable conclusions from complex datasets. The technique's adaptability to modern data challenges underscores its ongoing relevance in data analysis and visualization.

2.3.5 Stratified Sampling Technique

Stratified sampling is a sophisticated method used in statistics to increase the efficiency and accuracy of population analysis. This technique involves dividing a population into homogeneous subgroups, or strata, and then sampling from each stratum. It's particularly effective when different subgroups within a population vary significantly.

The theoretical foundation of stratified sampling is laid out in Cochran's "Sampling Techniques" [14]. It explains how stratified sampling can reduce sampling error compared to simple random sampling by ensuring that specific characteristics of the population are adequately represented.

Algorithmic Representation: The process of stratified sampling can be algorithmically represented as follows:

Algorithm 4 Stratified Sampling Technique

```
procedure StratifiedSampling(data, strata, sampleSize)
    Divide data into strata based on strata
    Initialize sampledData as an empty list
    for each stratum in strata do
        Determine sample size for stratum (proportional or equal)
        Perform random sampling within the stratum
        Append sampled data from stratum to sampledData
    end for
    return sampledData
end procedure
```

Stratified sampling is widely used in various research fields, including sociology, marketing, and environmental science. For example, in market research, stratified sampling allows businesses to understand preferences across different demographic groups, enhancing targeted marketing strategies.

In the era of big data, as discussed by Chen, Mao, and Liu in "Big Data: A Survey" [17], stratified sampling becomes crucial for analyzing large datasets efficiently. In data visualization, this technique allows for creating summary views that accurately represent different segments of a population, providing deeper insights into the data.

One of the primary benefits of stratified sampling is its ability to increase the precision of the results by reducing sampling error. Sedgwick's "Stratified Cluster Sampling" provides insights into how this method can be effectively implemented to ensure that all subgroups are adequately represented [15].

While stratified sampling is highly effective, it requires a thorough understanding of the population's characteristics to form appropriate strata. Incorrect stratification can lead to biased results, underscoring the need for careful planning and execution.

Lohr's "Sampling: Design and Analysis" emphasizes the need for a sound statistical approach in stratified sampling to ensure the reliability and validity of the results [6]. The method is particularly valuable in populations with diverse attributes, where simple random sampling might not provide an accurate representation.

Stratified sampling has evolved to address the complexities of modern datasets, including its application in areas like healthcare and social media analytics. Its adaptability and efficiency make it an invaluable tool in the toolkit of modern researchers dealing with heterogeneous populations.

Stratified sampling remains a vital and versatile technique in statistical analysis. Its ability to cater to diverse population attributes makes it indispensable in various research domains.

By ensuring representative and accurate sampling, stratified sampling continues to be a cornerstone in data analysis and visualization, particularly relevant in the context of large and diverse datasets.

2.3.6 Stratified Random Sampling

Stratified Random Sampling, an advanced form of sampling, enhances the representativeness and efficiency of statistical analysis. This method involves dividing a population into distinct subgroups or strata, and then conducting random sampling within each stratum. It combines the principles of stratification and randomness to ensure each subgroup is appropriately represented.

The concept and methodology of stratified random sampling are thoroughly explained in Cochran's "Sampling Techniques" [14]. This technique is ideal for populations with identifiable subgroups, ensuring that the sample includes elements from each segment proportionately.

Algorithmic Representation: The process of stratified random sampling can be algorithmically represented as follows:

Algorithm 5 Stratified Random Sampling Technique

```
procedure StratifiedRandomSampling(data, strata, sampleSizes)
    Divide data into strata based on strata
    Initialize sampledData as an empty list
    for each stratum in strata do
        n ← length(stratum)
        sampleSize ← sampleSizes[stratum]
        for i ← 1 to sampleSize do
            index ← RandomInteger(1, n)
            Append stratum[index] to sampledData
        end for
    end for
    return sampledData
end procedure
```

Stratified random sampling is widely used in areas like market research, healthcare studies, and social science research. In healthcare, for example, it allows researchers to ensure that various demographic groups (age, gender, etc.) are proportionately represented in studies.

In big data analytics, as discussed in "Big Data: A Survey" by Chen, Mao, and Liu [17], stratified random sampling is invaluable for analyzing large, diverse datasets. In data visualization, this method helps in creating summary views that accurately reflect the composition of different population segments, providing insights into each subgroup.

Stratified random sampling is particularly effective in reducing sampling bias and increasing the precision of the results. By ensuring that each stratum is represented, the technique minimizes the risk of overrepresentation or underrepresentation of any segment, as explained in Lohr's "Sampling: Design and Analysis" [6].

Implementing stratified random sampling requires in-depth knowledge of the population to form appropriate strata. Misclassification or improper stratification can lead to skewed results, emphasizing the need for accurate population understanding.

The reliability of stratified random sampling depends on the appropriate division of the population into strata and the random selection within each stratum. This method is particularly beneficial for heterogeneous populations, where simple random sampling might not capture the diversity effectively.

Stratified random sampling has evolved to meet the needs of modern data analysis, including applications in digital marketing and environmental studies. Its flexibility and efficiency make it a preferred choice for researchers dealing with complex and varied data sets.

Stratified random sampling continues to be a crucial technique in statistical analysis and data visualization. Its ability to cater to the unique attributes of diverse populations makes it essential in a wide range of research domains. By ensuring a balanced and accurate representation of subgroups, stratified random sampling remains indispensable in contemporary data analysis, especially in the context of large and heterogeneous datasets.

2.3.7 Systematic Sampling Technique

Systematic sampling is a widely used technique in statistical analysis where samples are selected at regular intervals from an ordered list. This method offers a blend of simplicity and efficiency, making it a popular choice for various types of surveys and research.

Systematic sampling's effectiveness and methodology are outlined in Cochran's "Sampling Techniques" [14]. The technique is particularly advantageous when dealing with large populations where a simple random sample might be too cumbersome to implement.

Algorithmic Representation: The process of systematic sampling can be algorithmically represented as follows:

Algorithm 6 Systematic Sampling Technique

```
procedure SystematicSampling(data, sampleInterval)
    n ← length(data)
    Initialize sampledData as an empty list
    startIndex ← RandomInteger(1, sampleInterval)
    for i ← startIndex to n step sampleInterval do
        Append data[i] to sampledData
    end for
    return sampledData
end procedure
```

Systematic sampling is utilized in numerous fields such as environmental studies, quality control, and market research. In environmental studies, for instance, it is used for sampling land areas or water bodies at regular intervals to assess pollution levels or biodiversity.

In the context of big data, systematic sampling is particularly useful for datasets with a large number of records. As highlighted in "Big Data: A Survey" by Chen, Mao, and Liu [17], it allows for efficient processing and analysis of large volumes of data. In data visualization, systematic sampling can help create representative visual summaries of large datasets, especially when the data is uniformly distributed.

While systematic sampling is efficient, it requires careful consideration of the interval selection to avoid biases, especially in datasets with hidden patterns. Lohr's "Sampling: Design and Analysis" discusses strategies to mitigate these risks and ensure the representativeness of the sample [6].

One challenge in systematic sampling is the risk of periodicity, where the sampling interval aligns with a pattern in the data, leading to biased results. Researchers must ensure that the sampling interval is not correlated with any pattern in the population to avoid this issue. The reliability of systematic sampling in statistical analysis is contingent on the random starting point and the absence of periodicity in the data. When executed correctly, it provides an efficient and effective method for sampling from large populations.

Systematic sampling has evolved to meet modern research requirements, including its application in digital analytics and healthcare research. Its efficiency and ease of implementation make it an attractive option for contemporary data analysis needs.

Systematic sampling remains a valuable tool in the arsenal of data analysts and researchers. Its utility spans a wide range of applications, from traditional surveys to modern big data analytics. By providing a balance between simplicity and effectiveness, systematic sampling continues to play a crucial role in extracting insights from large and varied datasets.

2.4 Summary of Sampling Techniques in Data Analysis

This chapter provided a comprehensive exploration of various sampling techniques, each vital for effective data analysis and visualization. From random to systematic sampling, these methods cater to diverse research needs, facilitating insightful analysis of large datasets.

Sampling techniques are crucial in data analysis, especially when dealing with expansive datasets. They enable the extraction of meaningful insights without requiring a full population study. The selection of an appropriate sampling method is fundamental to the accuracy and reliability of research outcomes, as emphasized in "Sampling: Design and Analysis" by Lohr [6].

Random and simple random sampling are foundational to statistical analysis, ensuring unbiased and representative sample selection. These methods are critical for the generalizability of research findings, as detailed in Anderson's "Introduction to Random Sampling" [5].

Stratified and cluster random sampling address the complexities of heterogeneous populations. By acknowledging distinct subgroup characteristics, these methods refine the precision of data analysis, making them particularly effective in diverse populations, as discussed in Cochran's "Sampling Techniques" [14].

Systematic sampling, known for its balance of simplicity and efficiency, is apt for uniformly distributed populations. It streamlines the sampling process while maintaining randomness, offering a practical solution for various research scenarios, as highlighted by Sedgwick in "Stratified Cluster Sampling" [15].

In data visualization, integrating these sampling techniques is key to creating representative and insightful visual summaries. They aid in developing visualizations that accurately reflect large datasets, as explored in "Big Data: A Survey" by Chen, Mao, and Liu [17]. This is crucial for understanding complex patterns and making data-driven decisions.

Sampling techniques, while beneficial, come with challenges like potential biases and implementation complexities. Future advancements in data analysis may focus on enhancing these methods, particularly for big data applications, to ensure more precise and insightful analyses, as suggested in "Visual Analytics: Definition, Process, and Challenges" by Keim et al. [21].

In summary, the sampling techniques discussed are indispensable in the realms of data analysis and visualization. Their proper application, as detailed in this chapter, is essential for researchers dealing with varied and large datasets. As data continues to grow in importance across multiple domains, the relevance of robust sampling methods in deriving meaningful insights remains critical.

Chapter 3

Enhancing Data Visualization and Summary Views for Large Datasets

3.1 Introduction

The advent of big data and advancements in computing technologies have transformed the landscape of data analysis, with data visualization emerging as a pivotal component. Chapter 3 delves into the multifaceted world of data visualization, particularly focusing on summary views for large datasets. This chapter is structured into two main parts: a review of existing work in the field and an exploration of innovative approaches developed in the project.

Data visualization, as a discipline, has evolved rapidly, transitioning from simple charting methods to complex, interactive visual analytics. This evolution is driven by the need to make sense of the ever-increasing volumes of data generated across various sectors. As we navigate through this chapter, we will explore how these methodologies have laid the groundwork for contemporary visualization techniques and how these approaches have been adapted to meet the challenges posed by large datasets.

One of the critical challenges in this domain is creating summary views that are both informative and manageable. Large datasets often contain a wealth of information that, if not visualized effectively, can lead to confusion or misinterpretation. This chapter will review traditional and current methods used to create summary views, discussing their strengths and limitations.

The second part of the chapter focuses on the project's contributions, where we have integrated sampling techniques with visualization tools to enhance the clarity and effectiveness

of summary views. This innovative approach aims to address some of the key challenges identified in earlier sections.

We will present the methodologies employed in the project, detailing the design and implementation of new visualization tools. A comparative analysis with existing methods will be provided to assess the effectiveness of the developed techniques. This section aims to highlight the project's contributions to the field of data visualization for large datasets.

As we conclude the chapter, we will reflect on the insights gained from the project and discuss potential future directions in data visualization. The rapidly evolving nature of data and technology presents both challenges and opportunities for innovation in this area.

Chapter 3 aims to provide a comprehensive overview of data visualization techniques for summarizing large datasets, from the perspectives to contemporary challenges and innovations. Through a blend of literature review and project-specific developments, this chapter contributes to the broader discourse on effectively visualizing complex data in an increasingly data-driven world.

3.1.1 Evolution of Data Visualization Techniques

The journey of data visualization dates back to when visual representation was used for mapping and recording data. Early efforts, such as John Snow's cholera map in 1854 [22], paved the way for the use of visuals in understanding complex information. As Tufte emphasizes in "The Visual Display of Quantitative Information," the ability to present data clearly and effectively is crucial [2]. This era laid the groundwork for future explorations in data visualization.

The 20th century witnessed significant advancements in statistical graphics, which were pivotal in transforming data visualization. Seminal figures like William Playfair, who is credited with introducing statistical graphs, and Florence Nightingale, known for her use of coxcomb plots in nursing, made substantial contributions. Their work, as highlighted in "Information Visualization: Perception for Design" by Ware [23], demonstrated the power of visual representation in interpreting data.

The advent of computers and digital technology marked a turning point in data visualization. As computing power increased, so did the ability to process large datasets and create more complex visualizations. This era saw the development of dynamic and interactive graphics, as discussed in "Interactive Data Visualization for the Web" by Murray [8]. The shift to digital opened new frontiers in visual analytics, blending data processing with graphical prowess.

William Playfair, considered a pioneer in statistical graphics, introduced the line chart, among other visual tools. His line chart from the 1786 book "The Commercial and Political Atlas" [24] represents the first time economic data was presented in a graphical form. It illustrates imports and exports between England and Denmark from 1700 to 1780, demonstrating a clear, visual method to interpret economic data trends over time. [25]

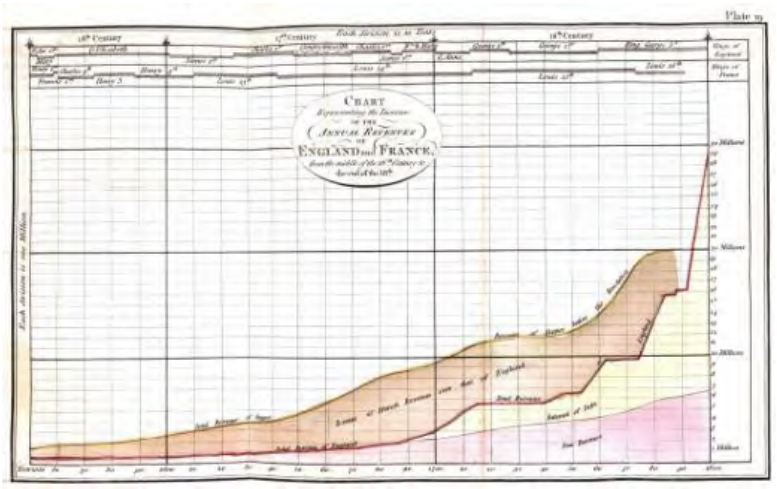

FIGURE 3.1: A chat that facilitates pre-attentive processing using different color to differentiate categories
[25]

The visualization here is an early example of William Playfair's work, a "Chart representing the Commercial History" [25] between England and Denmark and Norway. It exhibits the trade balance between these countries over 80 years, employing a line graph to depict imports and exports. This chart is significant for its introduction of area charts to represent data, marking a transformative point in the visual representation of statistical information. Playfair's innovative approach to economic data helped lay the foundation for modern statistical graphs, demonstrating the power of visualization in revealing trends and patterns within datasets.

The emergence of big data has significantly impacted data visualization. The challenge of making sense of vast amounts of information has led to innovative visualization techniques. Tools like D3.js have revolutionized the field, enabling the creation of sophisticated, interactive visualizations, as noted in "Beautiful Visualization: Looking at Data Through the Eyes of Experts" by Steele and Iliinsky [26].

The integration of machine learning and artificial intelligence in data visualization marks the current frontier in the field. These technologies offer ways to automate the creation of visualizations and extract insights from data more efficiently. As Johnson and Wichern discuss in "Applied Multivariate Statistical Analysis," AI and machine learning can uncover patterns and relationships in data that are not immediately apparent [27].

Recent trends emphasize the importance of user experience in data visualization. The focus has shifted towards making visualizations more accessible and intuitive. Works like "Visualization Analysis and Design" by Munzner [28] highlight the need to design visualizations that are not only informative but also engaging and easy to interpret for various audiences.

Looking forward, the field of data visualization faces both opportunities and challenges. As the volume and variety of data continue to grow, there is a need for more efficient and scalable visualization techniques. Emerging areas like augmented reality and virtual reality offer new possibilities for immersive data experiences, as suggested by recent research in the field.

The evolution of data visualization techniques has been marked by continuous innovation and adaptation to new challenges and technologies. From basic charting methods to advanced interactive and AI-driven visualizations, the field has undergone a remarkable transformation. As it progresses, the focus remains on making complex data comprehensible and actionable for a wide range of users.

3.1.2 Challenges in Visualizing Big Data

The advent of big data has brought forth unprecedented challenges in data visualization. Big data is characterized by its vast volume, high velocity, and varied formats, which pose significant hurdles in effectively visualizing information. As Chen, Mao, and Liu discuss in "Big Data: A Survey" [13], handling the sheer size and complexity of big data requires innovative visualization strategies that go beyond traditional methods.

One of the primary challenges in visualizing big data is scalability. The ability to process and visualize data in real-time or near-real-time is crucial, especially in domains where timely insights are vital. This issue is explored in "Interactive Dynamics for Visual Analysis" by Heer and Shneiderman [11], who emphasize the need for tools that can dynamically adapt to changing data flows and scales.

With the diversity of data sources, integrating heterogeneous data into a coherent visual format is a significant challenge. As Keim et al. note in "Visual Analytics: Definition, Process, and Challenges" [6], effective big data visualization requires harmonizing data from disparate sources, often involving complex data transformation and preprocessing steps.

Maintaining the accuracy and integrity of data during the visualization process is paramount. Misrepresentation or oversimplification of data can lead to incorrect conclusions. The work of Tufte in "The Visual Display of Quantitative Information" [2] highlights the ethical responsibility in accurately representing data, avoiding distortions and biases.

Creating user-friendly visualizations that cater to a diverse audience is another challenge. The goal is to design visual tools that are intuitive and accessible, even for users with limited technical expertise. Munzner's "Visualization Analysis and Design" [28] provides insights into creating effective visualizations that align with user needs and cognitive capabilities.

As datasets grow, there's a risk of information overload, where too much data can overwhelm the user. Strategies to simplify and summarize data without losing critical insights are essential. Techniques discussed in "Beautiful Visualization: Looking at Data Through the Eyes of Experts" by Steele and Iliinsky [26] offer ways to address this challenge by focusing on clarity and conciseness.

The use of advanced technologies like AI and machine learning in data visualization, as explored in "Applied Multivariate Statistical Analysis" by Johnson and Wichern [27], presents both opportunities and challenges. These technologies can automate and enhance visualization processes but require careful implementation to avoid misinterpretation of data.

Looking ahead, the field of data visualization is moving towards more interactive, immersive, and personalized experiences. The integration of augmented and virtual reality into data visualization, as indicated by recent research, opens new avenues for exploring and interacting with big data.

The challenges in visualizing big data are numerous and multifaceted, ranging from technical and computational issues to design and ethical considerations. Addressing these challenges requires a combination of advanced technologies, innovative design approaches, and a deep understanding of the underlying data and user requirements.

3.1.3 Role of Interactive and Dynamic Visualization in Big Data

In the realm of big data, interactive and dynamic visualization has emerged as a key tool for making complex data accessible and understandable. This approach allows users to explore and manipulate data visually, providing deeper insights and a more engaging experience. As Heer and Shneiderman discuss in "Interactive Dynamics for Visual Analysis," interactive visualization has transformed the way we interact with data [11].

The evolution from static charts to interactive graphics marks a significant shift in data visualization. Traditional static visualizations, while informative, limit the user's ability to explore data in depth. The advent of interactive tools, as highlighted in Murray's "Interactive Data Visualization for the Web" [8], enables users to drill down into specifics, discover patterns, and make data-driven decisions more effectively.

Gapminder World, developed by Hans Rosling with article of using Gapminder by Robert Lang [29], exemplifies interactive and dynamic visualization. It presents global development

data over time in an animated bubble chart, allowing users to observe trends and patterns across countries and indicators dynamically. This tool revolutionized the way big data is visualized and understood, emphasizing the importance of interactivity in data exploration.

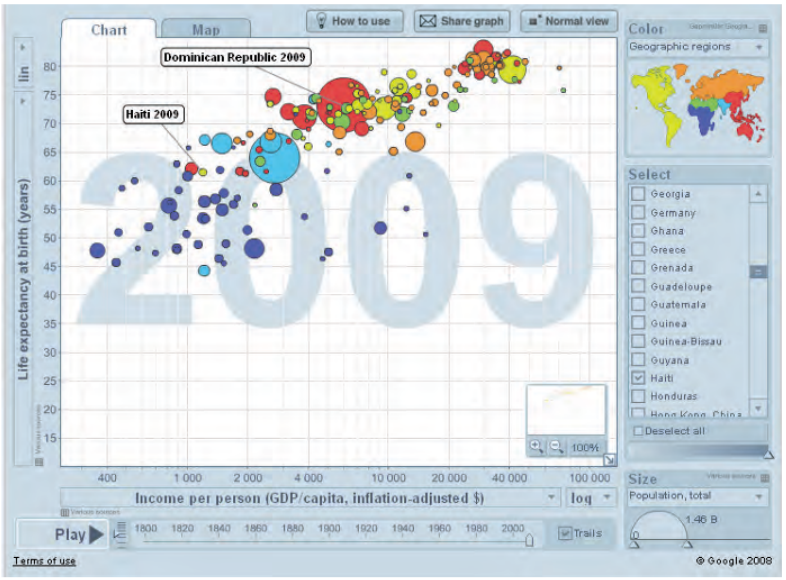

FIGURE 3.2: Gapminder screenshot looking at Haiti vs Domican Republic considering life expectancy and income
[29]

The image showcases a snapshot of the Gapminder World visualization tool [29], which is a dynamic and interactive scatter plot. This particular visualization displays data for the year 2009, depicting the relationship between life expectancy and income per person (GDP/capita, inflation-adjusted) for various countries. Each bubble represents a country, with its size corresponding to the country's population, and is colored based on geographic regions. Notable data points, such as the Dominican Republic and Haiti, are highlighted for comparison. Gapminder World's innovative approach allows for an engaging exploration of global health and economic trends over time, providing insights into how these two critical factors correlate across different nations and regions.

Interactive visualization enhances user engagement by allowing users to control their view of the data. This user-centric approach, explored in Munzner's "Visualization Analysis and Design" [28], ensures that visualizations are not only informative but also intuitive and tailored to the user's needs.

Advancements in technology, particularly in web-based applications and graphical interfaces, have propelled the growth of interactive visualization. Tools like D3.js have made it possible to create complex and responsive visuals that can handle large datasets efficiently, as discussed by Bostock et al. in "D3 Data-Driven Documents" [30].

Dynamic visualization addresses several challenges associated with big data, including scalability and real-time data processing. By leveraging technologies like machine learning and AI, as Johnson and Wichern describe in "Applied Multivariate Statistical Analysis" [27], dynamic visuals can adapt to changing data, offering insights that are both current and relevant.

Interactive and dynamic visualization has had a profound impact across various fields. In healthcare, for instance, it aids in understanding patient data trends, while in finance, it helps in analyzing market movements. These applications demonstrate the versatility and effectiveness of interactive visuals in deciphering complex data sets.

The future of interactive and dynamic visualization is poised for growth, with potential advancements in areas like augmented and virtual reality offering new ways to experience data. However, challenges such as data privacy, complexity management, and user accessibility remain pertinent, as indicated by recent research in the field.

Interactive and dynamic visualization plays a crucial role in the era of big data, offering enhanced user engagement, deeper insights, and a more intuitive understanding of complex datasets. As technology continues to evolve, so will the capabilities and applications of these visualization techniques, further solidifying their importance in data analysis.

3.1.4 Enhancing Summary Views with Big Data Techniques

The expansion of big data has necessitated the development of summary views, which are essential for distilling vast datasets into comprehensible visual formats. As outlined in "Big Data: A Survey" by Chen, Mao, and Liu [13], the challenge lies in extracting meaningful patterns and insights from large volumes of data while maintaining interpretability and accuracy.

Summary views are crucial in transforming raw data into actionable information. They allow users to grasp the essence of large datasets quickly, enabling more efficient decision-making processes. The work of Ware in "Information Visualization: Perception for Design" [23] emphasizes the importance of designing these views to cater to human perception and cognitive capabilities.

Recent technological advancements, particularly in data processing and visualization tools, have greatly enhanced the capability to generate effective summary views. The integration of machine learning algorithms, as explored in "Applied Multivariate Statistical Analysis" by Johnson and Wichern [27], has enabled the automated identification of relevant data patterns and trends.

Interactive and dynamic techniques have revolutionized summary views, offering users the ability to explore data in a more granular manner. As noted in Heer and Shneiderman's "Interactive Dynamics for Visual Analysis" [11], these techniques provide a user-centric approach, allowing for customizations and on-the-fly adjustments based on user interactions.

Creating effective summary views is not without challenges. Ensuring that these views are both representative of the underlying data and easily interpretable requires a careful balance. The potential for oversimplification or data misrepresentation is a constant concern, as Tufte discusses in "The Visual Display of Quantitative Information" [2].

The impact of enhanced summary views is evident across various domains, from healthcare, where they aid in patient data analysis, to finance, where they facilitate market trend analysis. The ability to quickly understand complex datasets has implications for policy-making, business strategies, and scientific research.

Looking ahead, the field of summary views in big data is poised for further innovations. The incorporation of augmented and virtual reality, as suggested by emerging research, could offer even more immersive and interactive data exploration experiences.

The enhancement of summary views with big data techniques represents a significant advancement in the field of data visualization. As technology continues to evolve, so will the methods and tools used to create these views, further empowering users to extract valuable insights from increasingly complex datasets.

3.1.5 Modern Tools and Technologies in Data Visualization

The advancement of data visualization tools and technologies has been integral in managing and interpreting the complexities of big data. Modern tools have evolved to offer more than just visual representation; they provide interactive, dynamic, and highly customizable experiences. This evolution is crucial, as noted in "Interactive Data Visualization for the Web" by Murray [8], in enabling users to engage with and understand complex datasets effectively.

Modern data visualization has benefited significantly from advancements in data processing software and visualization platforms. Tools like D3.js have become instrumental in creating interactive and responsive visuals, as described in Bostock et al.'s work on "D3 Data-Driven Documents" [30]. These tools allow for the manipulation of large datasets and the generation of real-time graphical representations.

Big data analytics tools have reshaped the landscape of data visualization. Software like Apache Hadoop and Spark provide the necessary infrastructure to process large volumes of

data, paving the way for more sophisticated visualization techniques. As Chen, Mao, and Liu highlight in "Big Data: A Survey" [13], these technologies are crucial for handling the volume, velocity, and variety characteristic of big data.

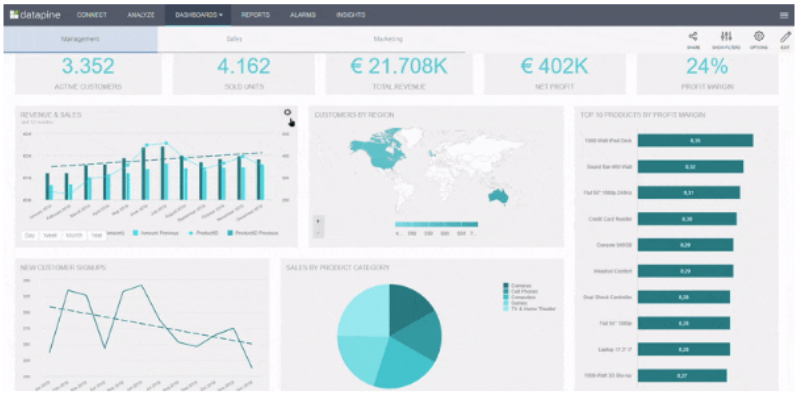

FIGURE 3.3: Comprehensive Business Dashboard: This image showcases an interactive dashboard that integrates various data visualizations, such as bar charts, line graphs, and maps, to convey critical business metrics at a glance.
[31]

The image depicts a multifaceted interactive dashboard designed to provide an at-a-glance overview of key business metrics. This type of dashboard is representative of modern data visualization tools that allow for real-time monitoring of diverse data streams. It incorporates a variety of visual elements such as bar graphs, line charts, pie charts, and maps, each designed to present specific types of data effectively. For instance, the bar graph shows revenue and sales trends over time, while the pie chart breaks down sales by product category, facilitating comparative analysis at a glance.

Such dashboards exemplify the advancements in data visualization technology by enabling users to interact with the data directly. The map is not merely a static display but a dynamic element that can provide regional customer data when interacted with. This interactivity enhances the user experience, making data exploration intuitive and accessible. The dashboard serves as an excellent example in this subsection, highlighting how modern visualization tools integrate complex data sets into cohesive and informative visual narratives, which are essential for data-driven decision-making in contemporary business environments.

The integration of artificial intelligence (AI) and machine learning into data visualization tools marks a significant leap forward. These technologies, as explored by Johnson and Wichern in "Applied Multivariate Statistical Analysis" [27], enable the automatic extraction of patterns and the generation of insights from large datasets, enhancing the effectiveness of visualization.

Cloud computing has also played a vital role in modern data visualization, offering scalable resources for processing and visualizing data. Collaborative tools and platforms have made it

possible for teams to work on data visualization projects remotely and in real-time, fostering a more collaborative and integrated approach.

The focus on user-centric design in modern visualization tools is pivotal. As Munzner discusses in "Visualization Analysis and Design" [28], designing visualizations that cater to the user's needs and cognitive capacities is essential for effective data interpretation and decision-making.

Despite these advancements, challenges remain, including ensuring data privacy, managing the complexity of visualization tools, and maintaining data integrity. The future of data visualization tools likely involves more immersive experiences, such as augmented and virtual reality applications, as suggested by emerging research in the field.

Modern tools and technologies in data visualization have transformed the way we interact with and interpret data. These advancements, from AI integration to cloud computing, have not only enhanced the capabilities of visualization tools but also expanded the possibilities for data exploration and understanding. As technology continues to evolve, the potential for innovation in data visualization remains vast and largely untapped.

3.2 Summary Views in Large Dataset Visualization

3.2.1 Traditional Methods for Summary Visualization

Before the advent of modern digital tools, traditional methods played a crucial role in summarizing and visualizing large datasets. These methods, rooted in statistical graphics and basic charting techniques, laid the foundation for today's advanced visualization technologies. Tufte's seminal work, "The Visual Display of Quantitative Information," highlights the importance of these early methods in effectively conveying data [2].

The evolution of traditional visualization techniques dates back to the use of simple line graphs, bar charts, and pie charts. These tools, while basic, were pivotal in representing data in a visual format. As described by Cleveland in "The Elements of Graphing Data" these techniques were essential for early data analysts to interpret and communicate their findings [32].

Statistical graphs played a significant role in early data analysis, offering a means to visually represent data trends and distributions. The development of histograms, scatter plots, and box plots provided analysts with tools to summarize large datasets effectively. Ware's "Information Visualization: Perception for Design" discusses the effectiveness of these graphical representations in enhancing data comprehension [23].

While traditional methods were groundbreaking for their time, they faced limitations, especially when dealing with large or complex datasets. Issues such as overplotting, data oversimplification, and lack of interactivity were common challenges, as noted in "Visualizing Data" by Cleveland [33].

The transition from manual graphing techniques to computer-aided visualization marked a significant shift in the field. This change, driven by the increasing complexity of datasets and the advent of computers, laid the groundwork for the sophisticated visualization tools we see today, as chronicled in "The Functional Art: An Introduction to Information Graphics and Visualization" by Cairo [34].

Comparing traditional methods with modern visualization techniques underscores the advancements made over the years. While traditional methods offered a solid foundation, modern tools provide enhanced capabilities such as interactive graphics, real-time data processing, and the ability to handle big data, as explored in "Interactive Dynamics for Visual Analysis" by Heer and Shneiderman [11].

Traditional methods for summary visualization were instrumental in the early days of data analysis. They paved the way for the development of more advanced, dynamic, and interactive visualization tools. Understanding these traditional methods provides valuable insights into the evolution of data visualization and underscores the ongoing need for effective data representation techniques.

3.2.2 Advances in Interactive Data Visualization

The evolution of data visualization has seen a significant shift towards interactivity, transforming how users engage with and interpret data. This transition is marked by the development of tools and platforms that allow for dynamic interaction with visualized data. As noted in "Interactive Data Visualization for the Web" by Murray [8], these advancements have revolutionized the field, enabling users to delve deeper into datasets and extract more nuanced insights.

The emergence of interactive visualization tools and platforms has been a game-changer in the field. Software like Tableau, QlikView, and D3.js have provided users with the ability to manipulate data visualizations in real-time, fostering a more engaging and exploratory experience. These tools, as discussed in Bostock et al.'s "D3 Data-Driven Documents" [10], allow for greater flexibility and creativity in data presentation.

Michael Bostock's D3.js library [35]enables the creation of data-driven documents, which has revolutionized the way we create interactive and dynamic visualizations on the web. An exemplary visualization could be a choropleth map that changes color based on data

inputs, allowing users to visualize complex geographical datasets interactively. This type of visualization could be used, for example, to represent population density or election results across different regions. The interactivity afforded by D3.js not only engages users but also allows for real-time data exploration and analysis, transforming static data into a rich, immersive experience.

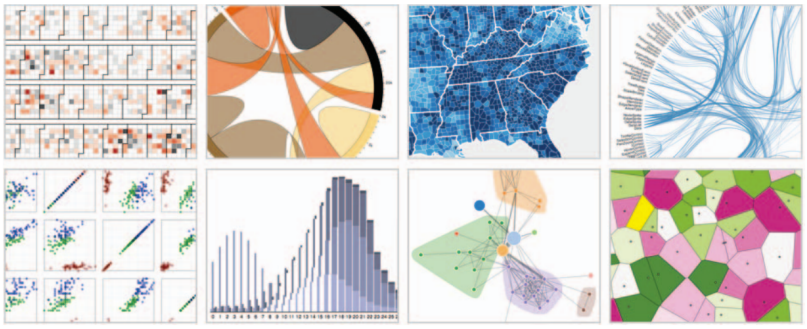

FIGURE 3.4: Interactive visualizations built with D3, running inside Google Chrome. From left to right: calendar view, chord diagram, choropleth map, hierarchical edge bundling, scatterplot matrix, grouped and stacked bars, force-directed graph clusters, Voronoi tessellation
[35]

The image presents a montage of complex and diverse visualizations, each representing different aspects and capabilities of modern data visualization techniques, likely created using D3.js as referenced in the work of Bostock et al [35]. From hierarchical treemaps to radial pie charts, choropleth maps, scatter plots, bar graphs, network diagrams, and Voronoi tessellations, each visualization offers a unique way to represent multidimensional datasets. These examples showcase the versatility of D3.js in translating raw data into comprehensible and interactive graphics that facilitate deeper insights and storytelling. Such visualizations exemplify the transformation of statistical figures into engaging visual narratives, allowing users to explore and interact with the data in intuitive ways. The collection emphasizes the creativity and technical expertise required to distill complex data into clear, impactful visual stories that can be understood at a glance by diverse audiences.

Interactive elements in data visualization, such as clickable icons, sliders, and hover effects, have greatly enhanced the user experience. They not only make data exploration more intuitive but also allow users to personalize their interaction with the data. The work of Heer and Shneiderman in "Interactive Dynamics for Visual Analysis" [11] highlights the significance of these elements in making complex data more accessible and understandable.

The rise of big data has further propelled the need for interactive visualization. With the increasing volume and complexity of data, interactive tools have become essential in managing and making sense of large datasets. Chen, Mao, and Liu, in "Big Data: A

Survey" [13], discuss how interactive visualization plays a critical role in big data analytics, enabling users to sift through large volumes of information efficiently.

The integration of machine learning and AI in interactive data visualization is an area of growing interest. As Johnson and Wichern articulate in "Applied Multivariate Statistical Analysis" [27], these technologies can enhance visualization tools by automating the extraction of patterns and facilitating predictive analytics.

Despite the advances, interactive data visualization faces challenges, including data privacy concerns, the steep learning curve of advanced tools, and the need for high computational power. Future directions, as suggested in "The Functional Art: An Introduction to Information Graphics and Visualization" by Cairo [34], may include further integration of AI and a focus on mobile-friendly and accessible designs.

The advances in interactive data visualization represent a significant stride in the field, offering users powerful tools to engage with data in meaningful ways. These developments have not only made data analysis more efficient but have also democratized access to complex data insights, paving the way for future innovations in the field.

3.3 Implementation of Sampling Techniques in Data Visualization for Large Datasets

Chapter 3 delves into the practical application of various sampling techniques in the realm of data visualization, particularly focusing on their implementation in a project aimed at summarizing and visualizing large datasets. This chapter is divided into two main sections: the first part reviews existing work and methodologies in data visualization, and the second part details the specific application and integration of these methods within our project.

The core of this chapter revolves around our project, which aims to enhance the visualization of large datasets by integrating different sampling techniques. This project represents an innovative approach to data visualization, where the primary challenge is not just to display data, but to do so in a manner that is both informative and manageable, even when dealing with vast amounts of information.

We explore how sampling techniques, discussed in detail in Chapter 2, can be effectively applied in data visualization. Our focus is on demonstrating the practical use of these techniques - namely random, stratified, systematic, and cluster sampling - in creating visual summaries of large datasets. The project aims to showcase how these sampling methods can enhance the clarity, accuracy, and interpretability of data visualizations.

The project also involves the development of various visualization tools and techniques, such as histograms, scatter plots, and other graphical representations. We discuss the design and implementation of these tools, emphasizing how they are specifically tailored to incorporate the sampling methods mentioned above.

A significant aspect of this chapter is the comparative analysis of our project's approach against traditional and contemporary data visualization methods. This comparison aims to highlight the effectiveness and efficiency of our integrated approach. Additionally, we examine user interaction and experience with our visualization tools, assessing their practicality and usability in real-world scenarios.

Chapter 3 aims to bridge the gap between theoretical understanding and practical application in the field of data visualization for large datasets. By detailing the implementation of sampling techniques in our project, this chapter contributes to the broader discourse on innovative approaches to data visualization, offering insights into both the challenges and successes encountered in our endeavor.

3.3.1 Integration of Random Sampling Technique in Visualization

In our data visualization project, we have integrated the random sampling technique as a pivotal method for summarizing and visualizing large datasets. Random sampling, a fundamental statistical method, is critical for managing the vast volume of data while ensuring representativeness and reducing computational load.

The Algorithm for Random Sampling in Visualization with integration of random sampling in our visualization tool is guided by a specific algorithm, designed to select a subset of data randomly for graphical representation. The algorithm is as follows:

Algorithm 7 Integrating Random Sampling in Data Visualization

1: **procedure** VisualizeWithRandomSampling(*data*, *sampleSize*, *visualizationFunction*)
2: $sampledData \leftarrow$ RandomSampling(*data*, *sampleSize*)
3: $visualization \leftarrow visualizationFunction(sampledData)$
4: **return** *visualization*
5: **end procedure**

In our project, as demonstrated in the Histogram.js file, random sampling is employed to generate histograms from large datasets. The sampling rate and the data subset are dynamically adjustable, allowing users to interactively explore different perspectives of the dataset. This approach, inspired by Heer and Shneiderman's principles in "Interactive Dynamics for Visual Analysis" [11], enhances the tool's flexibility and user engagement.

The use of random sampling in our visualization tool effectively reduces data complexity and computational requirements. This technique, while simple, is highly effective in maintaining the overall distribution and characteristics of the data, as discussed in "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman [36].

One challenge in implementing random sampling is ensuring the representativeness of the sample. To address this, we have incorporated mechanisms to adjust the sample size and to regenerate samples automatically through the algorithm defined, as suggested in "Sampling Techniques" by Cochran [4]. This ensures flexibility and accuracy in our visualizations.

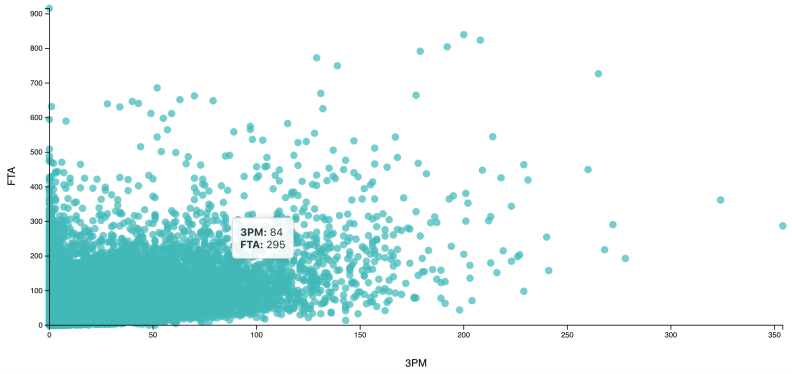

FIGURE 3.5: Displayed here is a random sampling of NBA data, with each point representing a correlation between 3-point shots made and free throws attempted. The plot emphasizes the uniform distribution of data points, reflecting the unbiased nature of random sampling.

The scatter plot above showcases the "Random Sampling" technique applied to a particular dataset, displaying 3-point shots made against free throws attempted. The randomness of the sampling is evident by the even distribution of data points across the graph. This technique does not show any specific pattern or clustering, which underscores the randomness of the selection process.

Random sampling is generated by selecting a subset of individuals from the larger dataset where each individual has an equal chance of being chosen. In this dataset, each player or game statistic has an equal likelihood of inclusion in the sample. This method ensures that the sample is unbiased and that the results are generalizable to the larger population.

The integration of random sampling has significantly enhanced the user experience. Users can interact with the visualization, experimenting with different sample sizes and observing the changes in the graphical output by selecting number of bins for instance for a paticular given dataset. This interactive feature aligns with the user-centric design principles in data visualization, highlighted in Munzner's "Visualization Analysis and Design" [28].

The integration of random sampling in our data visualization project represents a strategic approach to managing and interpreting large datasets. By effectively balancing data representativeness with computational efficiency, this technique has proven crucial in enhancing our project's visualization capabilities.

3.3.2 Integration of Stratified Sampling Technique in Visualization

In our project, we implemented stratified random sampling to enhance the accuracy and relevance of our data visualizations. This technique, vital for handling diverse and heterogeneous datasets, ensures that each distinct subgroup within the data is appropriately represented in the visual output.

Algorithm for Stratified Sampling in Visualization with the process of integrating stratified random sampling into our visualization tool is governed by a specific algorithm, as follows:

Algorithm 8 Integrating Stratified Random Sampling in Data Visualization

procedure VisualizeWithStratifiedRandomSampling(*data*, *strataCriteria*, *sampleSizePerStratum*, *visualizationFunction*)
    Divide *data* into strata based on *strataCriteria*
    Initialize *sampledData* as an empty list
    for each *stratum* in *data* **do**
        stratumSample ← SimpleRandomSampling(*stratum*, *sampleSizePerStratum*)
        Append *stratumSample* to *sampledData*
    end for
    visualization ← *visualizationFunction*(*sampledData*)
    return *visualization*
end procedure

In our project, stratified sampling is applied to ensure diverse data segments, such as different player categories or performance levels, are equitably represented in the visualizations. This approach aligns with the principles advocated by Cochran in "Sampling Techniques" [4], enhancing the analytical accuracy of our visualizations.

The use of stratified random sampling has significantly improved the accuracy and relevance of the visualizations produced in the project. By ensuring that each stratum is properly represented, our visualizations offer a true reflection of the entire dataset, capturing its diversity and intricacies.

Stratified sampling is particularly effective in addressing the challenges posed by the heterogeneous nature of large datasets. It simplifies data complexity by breaking it down into more homogeneous subgroups, making it more manageable for visualization and analysis.

The above scatter plot with stratified sampling technique applied demonstrates the data for 3-point shots made versus free throws attempted, segmented into distinct strata or groups.

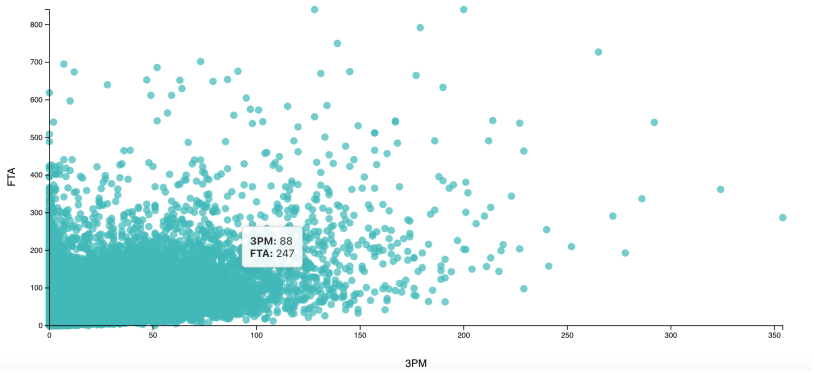

FIGURE 3.6: This graph depicts stratified sampling of NBA statistics, segregating data into distinct strata before sampling. It demonstrates the variance within each stratum, ensuring each category is proportionally represented

The stratification is typically based on a key variable or category that is important to the research question, which might be player positions or team divisions in this case

Stratified sampling involves dividing the population into homogeneous subgroups or strata and then taking a random sample from each stratum. For the dataset, the population could be divided by position (guards, forwards, centers), and then a random sample from each group is used to form the final sample. This technique ensures that each subgroup is adequately represented in the sample, which can provide more precise estimates than simple random sampling.

Integrating stratified random sampling in our data visualization project has enriched the analytical process. It ensures that the visualizations are not only representative of the entire dataset but also tailored to highlight specific characteristics within the data. This technique, combined with our visualization tools, offers a nuanced understanding of complex datasets.

3.3.3 Integration of Systematic Sampling Technique in Visualization

Systematic sampling, a variant of random sampling, is employed in this project for its efficiency in handling large datasets. It involves selecting samples at regular intervals from an ordered population, blending randomness and structure. This method is particularly suitable for time-sequenced data or datasets with inherent order, such as analyzing player performance over a season.

Algorithm for Systematic Sampling with the systematic sampling algorithm used in the project is as follows:

Systematic sampling's appeal lies in its balance between ease of implementation and the ability to produce representative samples. It is advantageous when the population exhibits

Algorithm 9 Integrating Systematic Sampling in Data Visualization

```
procedure VisualizeWithSystematicSampling(data, interval, visualizationFunction)
    n ← length(data)
    Initialize sampledData as an empty list
    start ← RandomInteger(1, interval)
    for i ← start to n by interval do
        Append data[i] to sampledData
    end for
    visualization ← visualizationFunction(sampledData)
    return visualization
end procedure
```

a natural ordering. The theoretical underpinning of this method is supported by Lohr's "Sampling: Design and Analysis" [7].

Systematic sampling is highly effective for temporal datasets, like tracking performances across seasons or tournaments. It facilitates consistent data selection at periodic intervals, aiding in chronological analysis to uncover significant trends.

While less random than simple or stratified random sampling, systematic sampling maintains a degree of randomness through its initial random start, ensuring the sample is not biased by positional patterns in the data, as highlighted by Cochran in "Sampling Techniques" [4].

One challenge is the risk of periodicity bias if the interval aligns with a pattern in the data. To mitigate this, the project involves careful data examination beforehand to choose an appropriate interval as per the algorithm of selecting random integer, minimizing this risk as recommended by Sedgwick in "Systematic Sampling" [10].

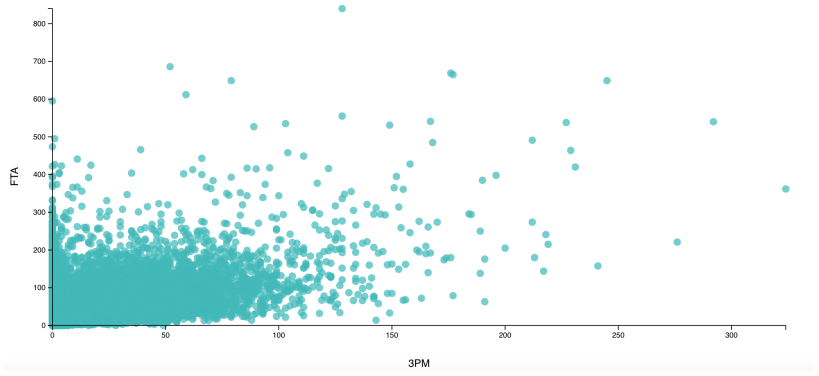

FIGURE 3.7: Presented is the systematic sampling technique, where data points are selected at regular intervals from the NBA dataset. This plot shows a methodical selection pattern across the 3PM and FTA variables

In the "Systematic Sampling" scatter plot, the data points of 3-point shots made versus free throws attempted are spread in a pattern that appears more regular and systematized than

in simple random sampling. This may suggest a fixed interval of selection, such as every nth player or game in the dataset.

Systematic sampling is conducted by selecting data points from the population at a regular interval, determined by a starting point and a fixed periodic interval. In the dataset context, if the dataset is ordered in some way (e.g., by date of game or player's alphabetical listing), every nth entry could be selected for inclusion in the sample. This method is often used when an ordered list of the population is available and can often be easier and quicker to carry out than random sampling.

In summary, systematic sampling is an efficient and effective method for data analysis in this project. Its application bolsters the ability to capture and visualize key trends, significantly contributing to the project's data analysis framework.

3.3.4 Integration of Cluster Sampling Technique in Visulization

In our data visualization project, we leverage cluster random sampling for its efficiency in handling extensive datasets that are naturally segmented, such as those categorized by leagues, age groups, or geographical regions. This method involves analyzing representative clusters to gain insights reflective of the broader population while minimizing computational load.

Algorithm for Cluster Sampling with the implementation of cluster random sampling in our visualization tool is based on the following algorithm:

Algorithm 10 Integrating Cluster Random Sampling in Data Visualization

procedure VisualizeWithClusterSampling(*data*, *numClusters*, *visualizationFunction*)
    Divide *data* into clusters based on predefined criteria
    Select *numClusters* randomly
    Initialize *sampledData* as an empty list
    for each *cluster* in selected clusters **do**
        Append all data points from *cluster* to *sampledData*
    end for
    $visualization \leftarrow visualizationFunction(sampledData)$
    return *visualization*
end procedure

Cluster random sampling is founded on the principle of capturing population characteristics by analyzing subsets (clusters) of the population. This approach, rooted in the works of renowned statisticians, offers a strategic method for visualizing complex data structures.

This sampling method is applied in our project to efficiently and effectively analyze data that is segmented into natural groups. By sampling entire clusters, the project ensures that each segment is adequately represented in the analysis, providing a holistic view of the data.

Cluster random sampling is particularly effective for dealing with large, geographically dispersed datasets. It simplifies data complexity by focusing on clusters, thereby making the data more manageable for visualization and analysis.

A key consideration in implementing this method is ensuring that the clusters themselves are representative of the overall population. Our project addresses this by carefully defining criteria for cluster formation, ensuring that each cluster is as heterogeneous as the overall population.

The scatter plot in the figure depicts the relationship between 3-point shots made (3PM) and free throws attempted (FTA) in an NBA dataset. It shows a spread of data points across the chart, with concentrations in certain areas, suggesting the clusters from which the samples were drawn. The visualization highlights areas with a higher density of data points, as well as outliers.

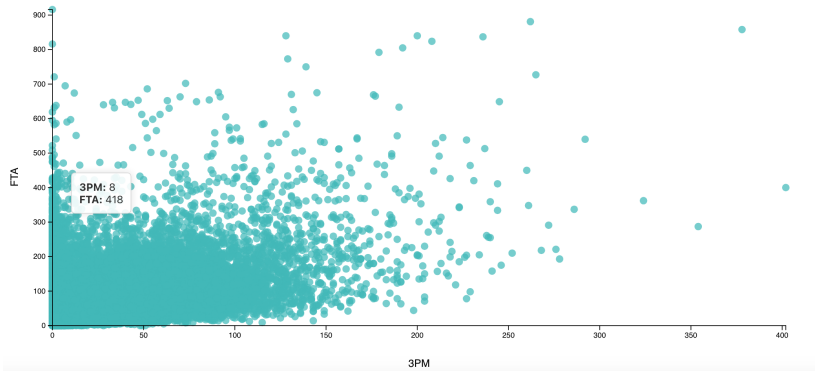

FIGURE 3.8: This visualization illustrates the cluster sampling method applied to given dataset, highlighting the relationship between 3-point shots made (3PM) and free throws attempted (FTA). The clusters represent grouped data points selected to provide a comprehensive overview of the dataset's distribution.

Cluster sampling involves dividing the dataset into clusters that are then randomly selected. In the context of the given dataset, players or games could be grouped into clusters based on similar characteristics (like teams or positions). A number of these clusters are then chosen at random, and data from all members of these clusters are used to create the visualization. This method is particularly useful when dealing with large datasets, as it provides a manageable, yet representative, subset of data.

In conclusion, the integration of cluster random sampling into our data visualization tools is a testament to the project's innovative approach to data analysis. By leveraging this technique, the project not only enhances the efficiency of its data handling but also ensures

that the visualizations are rich, informative, and reflective of the complex nature of the datasets at hand.

3.3.5 Effective Use of Sampling in Summary Views

In our visualization project, we effectively utilized various sampling techniques to generate summary views of large datasets. These techniques are instrumental in managing the complexity and size of data, ensuring that our visualizations are both efficient and representative of the underlying information.

We integrated multiple sampling methods, including random, stratified, and systematic sampling, to cater to different data types and visualization requirements. This integration allowed us to balance the need for detailed analysis with the constraints of processing large datasets.

Random sampling was particularly effective in dynamic data analysis scenarios, providing real-time generation of visualizations that could adapt to different data segments or timeframes. This flexibility was crucial for interactive data exploration.

Stratified random sampling was used to ensure that visualizations were comprehensive and insightful, especially when dealing with multifaceted data like player performances across different positions or leagues. This method enhanced the precision of our visualizations by focusing on specific subgroups within the data.

Systematic sampling proved to be a reliable approach for sequential data analysis, enabling us to visualize data trends over time or across structured datasets. This technique ensured a comprehensive yet concise representation of the data, maintaining data integrity while delivering efficient visualizations.

Our application of sampling techniques was guided by a user-centric approach, allowing users to interactively explore data subsets, change visualization types, and focus on areas of interest with ease. This adaptability exemplified the power of merging visualization with sampling and sorting algorithms.

The integration of these techniques ensured that our application was adaptable to user preferences, offering a spectrum of visualization options. Whether it's tracking player performance or analyzing historical statistics, our application maintained responsiveness and relevance.

In summary, the effective use of sampling techniques in our project exemplified best practices in data analysis. By integrating these techniques with data visualization tools, we managed

to provide insightful and interactive experiences, navigating the challenges of large dataset visualization with finesse.

3.4 Visualization Tools and Techniques

This section of Chapter 3 focuses on the visualization tools and techniques employed in our project to effectively present and interpret large datasets. Our approach centers on the integration of advanced sampling methods with innovative visualization tools, tailoring the visualization process to meet the challenges of big data.

Our project underscores the importance of interactive and dynamic visualization tools. This emphasis aligns with the principles discussed in "Interactive Data Visualization for the Web" by Murray [8], which highlights the necessity of engaging and flexible visualization techniques in the era of big data.

The visualization tools developed in our project are customized to effectively utilize different sampling techniques, such as random, stratified, and systematic sampling. These tools are designed to handle various types of data, offering users a range of visualization options like histograms and scatter plots.

A significant aspect of our visualization approach is the seamless integration of sampling methods into the tools. This integration, as outlined in "The Visual Display of Quantitative Information" by Tufte [2], ensures that the visualizations are not only representative of the larger dataset but also maintain clarity and precision.

Our visualization tools are grounded in a user-centric design philosophy. We prioritize ease of use and intuitive interfaces, as emphasized in Munzner's "Visualization Analysis and Design" [28]. This approach ensures that users can effectively interact with and interpret the visualized data, regardless of their technical background.

One of the key challenges addressed in our project is balancing the complexity of data with the usability of visualization tools. We aim to provide users with the capability to delve into complex datasets without being overwhelmed, a concept explored in "The Functional Art: An Introduction to Information Graphics and Visualization" by Cairo [34].

In conclusion, this section outlines our project's approach to developing and implementing visualization tools and techniques. By combining innovative sampling methods with dynamic and user-friendly visualization tools, our project offers a sophisticated solution to the challenges of visualizing large datasets.

3.4.1 Design and Development of Visualization Tools

In our data visualization project, we focused on designing and developing tools that effectively integrate various sampling methods to visualize large datasets. This section delves into the core visualization tools used in the project, namely histograms and scatter plots, and discusses their development process.

Our project utilized histograms and scatter plots as primary tools for data visualization. Histograms were instrumental in showcasing frequency distributions, particularly useful in understanding the spread and central tendencies within the data. Scatter plots, on the other hand, provided insights into the relationships between different data variables.

We began by identifying the key requirements for our visualization tools, focusing on usability, scalability, and the ability to incorporate different sampling techniques. We designed the tools with a user-friendly interface, allowing for interactive exploration of data. This design process was informed by principles from Tufte's "The Visual Display of Quantitative Information" [2]. The tools were developed using modern web technologies, ensuring they are accessible and responsive. The integration of D3.js, as explored in Murray's "Interactive Data Visualization for the Web" [8], played a crucial role in this phase.

A key feature of our tools is the integration of sampling methods. For example, the histogram tool could dynamically change its presentation based on the selected sampling technique, whether it was random, stratified, or systematic sampling. Similarly, the scatter plot tool could visualize data points sampled using various methods, offering a versatile view of the dataset.

Throughout the development process, we faced challenges related to data handling and performance optimization. We addressed these by implementing efficient data processing algorithms and optimizing the tools for large datasets, as suggested in "Visualization Analysis and Design" by Munzner [28].

User experience was a focal point in our development process. We continuously sought feedback from users to refine the tools, ensuring they met the needs of data analysts and researchers. The interactive and dynamic nature of the tools, as highlighted in "The Functional Art: An Introduction to Information Graphics and Visualization" by Cairo [34], greatly enhanced user engagement.

In summary, the design and development of visualization tools in our project were guided by the goal of effectively integrating sampling techniques into data visualization. The resulting tools, namely histograms and scatter plots, provide robust, interactive, and insightful visualizations of large datasets, demonstrating the project's commitment to innovative data analysis.

3.4.2 Custom Implementation of Visualization Techniques

Our project's data visualization tools, such as the HistogramChart.js, Chart.js, and ScatterPlot.js modules, embody the intricate relationship between data processing and visualization. The custom implementation of these tools plays a pivotal role in presenting data subsets adaptively, based on user interactions and preferences.

The modules were designed to adaptively select and present data subsets, a critical feature in managing large datasets. This adaptability ensures the visualizations remain responsive and informative, even when handling extensive data. Sampling techniques like random, systematic, and stratified sampling underpin this dynamic visualization process.

For instance, the HistogramChart.js component employs random sampling to enhance the agility of histogram chart visualizations. This approach strikes a balance between data reduction and trend preservation, boosting chart readability and expediting rendering, which is essential for large datasets.

The design philosophy of our visualization tools was to create efficient, powerful tools for data analysis, ensuring they seamlessly interact with the data processing backend. This approach aligns with the principles outlined in "Visual framework for big data in D3.js" by Bao and Chen [37], which emphasize the importance of modular and reusable user interface elements in crafting effective visual narratives.

One of the significant challenges in the project was handling the computational demands of large datasets. Our custom implementation focused on achieving a delicate balance between meaningful data representation and computational feasibility. Robustness against outliers was also a focal point, ensuring that the outcomes were not unduly influenced by exceptional cases.

In conclusion, the custom implementation of visualization techniques in our project was a crucial aspect of its success. By leveraging modern web technologies and innovative data processing strategies, we ensured that our visualizations were not only accurate and informative but also responsive and user-friendly. This approach highlights the project's commitment to advancing the field of data visualization through innovative techniques and user-centered design.

3.4.3 Features providing Comprehensive Summary Views

Our data visualization project is tailored to provide comprehensive views of large datasets by integrating various visualization techniques and features. This approach enhances the

user's ability to gain insights from complex data, aligning with the objectives outlined in "Data Visualization: A Practical Introduction" by Healy [9].

Central to our visualization suite are dynamic histograms and scatter plots, which adjust in real-time based on user interactions and sampling choices. This flexibility is crucial for exploring different aspects of the data effectively. The Chart module allows users to trace data trends over time, offering an interactive and temporal perspective. This tool is particularly useful for datasets with chronological elements, providing a clear evolution of data points.

NASA's "Eyes on Asteroids" [38] visualization is another excellent example of a tool that provides summary view for comprehensive data. This interactive tool allows users to explore the asteroid belt and see the real-time positions of asteroids in our solar system. It features a 3D model of our solar system, with a sleek and modern design that provides an engaging user experience. The real-time data feed on asteroid positions and trajectories makes it highly relevant and informative. "Eyes on Asteroids" is a prime example of how interactivity can enhance the user experience in data visualization, making complex space data accessible and understandable to the general public.

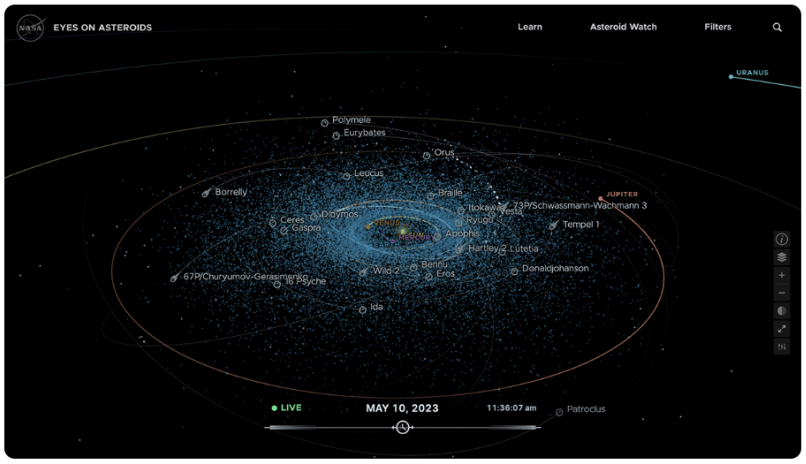

FIGURE 3.9: NASA's 'Eyes on Asteroids' visualization showcases the real-time tracking of asteroids in our solar system. The detailed 3D model provides an engaging and educational view of asteroid positions and trajectories as of May 10, 2023, offering an insightful glimpse into our dynamic cosmos.
[38]

This tool uses real-time data and 3D modeling to provide an up-to-date and interactive exploration of celestial bodies. Users can zoom in and out of the solar system, making it an effective educational tool for understanding the dynamics of asteroids in relation to Earth. The visualization "Eyes on Asteroids" [38] is a remarkable representation of NASA's asteroid tracking system, providing a real-time 3D model of our solar system and the asteroids within it. As of May 10, 2023, this interactive tool shows the scattered distribution of asteroids

with trajectories that illustrate their orbits around the sun, bringing attention to celestial bodies such as Ceres, Apophis, and many others. The user interface includes features for learning, asteroid watching, and applying various filters, enhancing the educational value and user engagement with the cosmic environment beyond Earth. This visualization is not only a feat of technical and design prowess but also serves as a crucial tool for both public interest and scientific communities to monitor and study near-Earth objects.

A significant aspect of our project is the 'Summary View' feature, which allows users to view the entire dataset or a comprehensively sampled subset into a summarized view. This feature is instrumental in giving a holistic view of the data, ensuring that no critical information is overlooked. Users can customize various parameters, such as sample size, data range by selecting and changing the number of bins, and also visualization type. This customization, as suggested in "The Functional Art: An Introduction to Information Graphics and Visualization" by Cairo [34], empowers users to tailor the visualization experience to their specific needs.

Through the effective use of labeling, and axis scaling, our visualizations offer enhanced data representation. This approach aligns with Tufte's principles in "The Visual Display of Quantitative Information" [2], emphasizing clarity and precision in data visualization. The project integrates advanced sampling techniques like stratified and cluster sampling in the visualization tools. This integration ensures that the visualizations are not only comprehensive but also representative of the entire dataset.

One of the challenges in visualizing large datasets is maintaining performance and readability. Our project addresses this by optimizing data processing algorithms and visualization rendering, ensuring that the tools are both efficient and user-friendly.

In summary, the comprehensive view features in our project represent a blend of innovative visualization techniques and user-centered design. By offering dynamic, interactive, and customizable tools, our project facilitates an in-depth exploration of large datasets, ensuring that users can derive meaningful insights from complex data visualizations.

Chapter 4

Visualization Tools and Techniques

4.1 Introduction

The realm of data visualization has transformed the way we interpret and engage with data. Effective visualization acts as a bridge between the raw data and its comprehensive understanding, emphasizing the need for skilled techniques and tools in this field [1]. This chapter delves into the specifics of utilizing D3.js, React, and JavaScript, pivotal tools in the modern landscape of data visualization [8].

Today's data-driven world demands robust visualization tools to effectively communicate complex information. The art and science of visualizing data not only make data more accessible but also uncover hidden patterns and insights. This becomes particularly essential in big data contexts, where traditional analysis methods falter.

This chapter will explore D3.js, a powerful JavaScript library for creating dynamic, interactive data visualizations in web browsers [8]. The integration of React, a JavaScript library for building user interfaces, with D3.js, enables the development of complex, reactive visualizations that are both informative and engaging. The synergy of these tools exemplifies the evolution of data visualization practices, aligning with the trends and demands of current data analytics [11].

The choice of these tools for the project was influenced by their versatility, robustness, and wide acceptance in the data visualization community. React's component-based architecture complements D3's data-driven approach, offering a scalable solution for complex visualizations. This chapter will demonstrate how these tools were employed to address specific project requirements, enhancing the visualization capabilities of the project and providing a richer, more interactive user experience [14].

4.1.1 Overview of Visualization Tools and Techniques

Data visualization has undergone significant evolution, growing from simple charting techniques to complex, interactive visualizations that offer deep insights into large datasets [1]. This transformation has been driven by the increasing complexity of data and the need for more sophisticated tools to understand and communicate this data effectively.

D3.js stands out as a versatile tool in data visualization, offering flexibility and control over the final visual output. Its ability to bind data to a Document Object Model (DOM) and apply data-driven transformations to the document makes it a powerful tool in the visualization toolkit [30].

React's introduction has revolutionized the way web applications are built, particularly in the realm of interactive user interfaces [19]. Coupled with JavaScript, it enables the creation of responsive and dynamic visualizations, which are crucial for engaging and informative data presentations.

This chapter aims to provide a comprehensive understanding of how D3.js, React, and JavaScript were utilized in the project to create effective visualizations. It will explore the specific features of these tools that make them suitable for various visualization tasks in the project [19].

4.1.2 Objectives

The primary objective of this chapter is to deepen the understanding of data visualization's role in modern data analysis. With the increasing complexity of data, visualization has become an indispensable tool in making data comprehensible and actionable [1].

A key focus is to explore the capabilities of D3.js in creating dynamic and interactive data visualizations. D3.js's ability to bind arbitrary data to the Document Object Model (DOM) and apply data-driven transformations provides unparalleled flexibility.

Another objective is to examine how React, combined with JavaScript, enhances the interactivity and responsiveness of data visualizations. This combination allows for the creation of complex user interfaces that are both functional and visually appealing. The chapter aims to demonstrate the practical application of these tools within the context of the project. This involves showcasing specific instances where D3.js, React, and JavaScript were utilized to solve real-world visualization challenges [30].

The chapter will also engage in a comparative analysis, contrasting the chosen tools and techniques with other available options in the field. This analysis will underline the reasons behind the selection of these specific tools for the project. An essential objective is

to highlight the challenges encountered during the implementation of these tools and the solutions that were devised. This will provide insights into the practical aspects of working with complex visualization tools [14].

Finally, the chapter aims to bridge the gap between theoretical knowledge and practical application. By correlating academic principles with real-world implementation, it aims to provide a comprehensive understanding of the current state of data visualization.

4.2 Introduction to the Visualization Tool

In Section 4.2, we unveil the innovative visualization tool developed in this study, designed to enable intricate data analysis through a suite of dynamic and interactive visualizations. Building upon the foundational principles outlined in Chapter 3 , this tool embodies the convergence of theoretical data visualization techniques and practical application. The visualization tool, a culmination of the technologies discussed in Sections 4.4 and 4.5, is specifically tailored to address the challenges of interpreting complex datasets, as characterized by the diverse sampling methods highlighted in Chapter 3.

The user-centric interface facilitates an intuitive exploration of data, allowing for both granular and macroscopic insights, while the backend architecture ensures robust performance even with substantial data loads. This section guides the reader through the tool's features, its operational workflow, and the tangible benefits it delivers in the realm of data visualization.

4.2.1 Tool Overview

In this subsection, we delve into the visualization tool designed to harness the power of data through interactive and dynamic visuals. Rooted in the methodologies discussed in Anderson's "Introduction to Random Sampling" [3], the tool leverages the robust frameworks of D3.js similar to Murray's "Interactive data visualization for the web" [8] and React to create a suite of visualizations that represent data. We will explore the design philosophy that emphasizes simplicity in interaction while handling complex data sets, as highlighted by Heer and Shneiderman's "Interactive dynamics for visual analysis" [11].

The tool's ability to represent multi-dimensional datasets through varied visualization techniques, such as scatter plots and histograms, will be showcased. These techniques are inspired by the foundational work of Tufte's "Visual Display of Quantitative Information" [2] and William's "Elements of graphing data" [32], ensuring that each visual element conveys information effectively and efficiently. The integration of React enhances the reactivity of

the visualizations, making the tool dynamic and responsive to user interactions similar to Vo and Lam's work of "Web application development with react" [19].

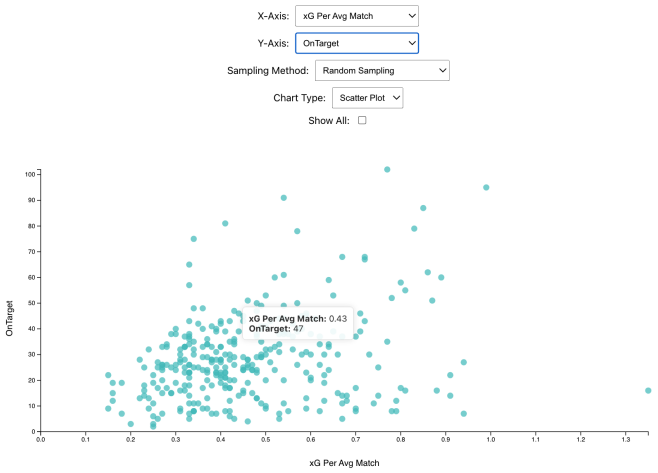

FIGURE 4.1: This visualization from the data analysis tool demonstrates the interactive capabilities, showing the relationship between 'xG Per Avg Match' and 'OnTarget' in a football dataset

The interactive visualization presented here is a fundamental component of the data analysis tool detailed in this project. It showcases a scatter plot that dynamically represents the relationship between expected goals per average match (xG Per Avg Match) and shots on target, illustrating the tool's capacity to translate complex data into accessible insights. The image embodies the tool's interactive nature, allowing users to adjust variables and observe real-time changes in data relationships, a core feature highlighted in the tool's overview.

This scatter plot exemplifies how the tool operationalizes data visualization theories and practices as discussed in Iliinsky's "Beautiful visualization: Looking at data through the eyes of experts" [26], reinforcing the narrative that visualization is not merely a presentation technique but an analytical process. The image serves as an overview of the tool's functionality, reflecting on the integration of user interactivity with data-driven decision-making, ensuring that users are not passive observers but active participants in the data exploration journey.

The paragraph about the image will explain how the histogram is generated using a random sampling method based on expected goals per average match played to provide a balanced view of the dataset, drawing from sampling techniques. This visualization exemplifies the tool's capability to provide insights into large datasets, enabling users to discern patterns and outliers in the data swiftly.

In summary, this section will provide a comprehensive overview of the visualization tool's capabilities, from data processing to user interaction, ensuring a clear understanding of its utility and effectiveness in the field of data visualization.

4.2.2 Visualization Features

This subsection showcases the core visualization capabilities of our tool, which is central to our project. Each visualization is not just a static image but a comprehensive summary view of the dataset, allowing for an analysis of the data through various sampling techniques. This section will introduce the main visualization features our tool offers, the technical aspects of how these visualizations are generated, and their relevance in the context of data analysis.

Starting with an overview of the tool's visualization features, we delve into the implementation of histograms, scatter plots, and line charts, each offering a unique lens through which to view the data. These visualizations are rooted in established data visualization practices, as discussed by Cleveland in "The Elements of Graphing Data" [32] and Tufte in "The Visual Display of Quantitative Information" [2].

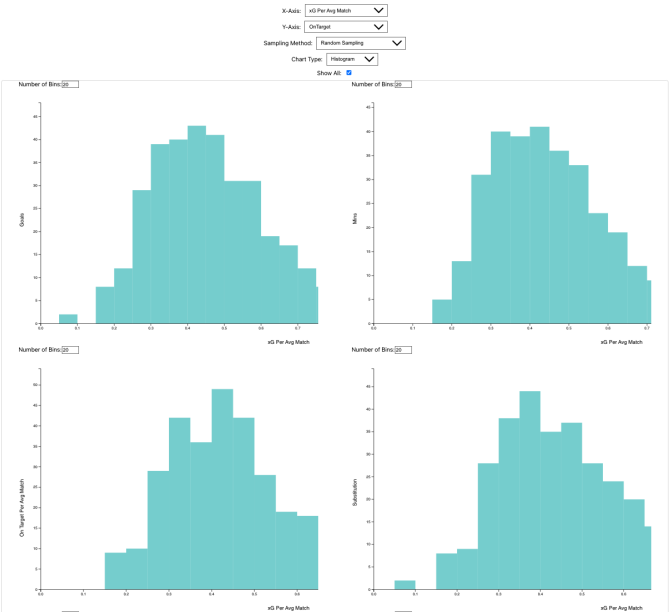

FIGURE 4.2: The tool generates multiple histogram visualizations using random sampling, offering a comprehensive snapshot of the dataset across various metrics

As shown in image, it exemplifies one of the tool's key features: the generation of multiple histograms through random sampling to provide a sampled summary view. This feature is particularly useful for large datasets where an overview is required before diving into a detailed analysis. It reflects the principles outlined in "Data Visualization in Sociology" [1] and "A survey of data partitioning and sampling methods to support big data analysis" [5].

The histograms depicted show the distribution of 'xG Per Avg Match' across different subsets of the data, with the Y-axis representing the count of observations in each bin. This visual aggregation helps in quickly identifying patterns, such as central tendencies and skewness,

within the dataset. The visualization technique is informed by Healy and Moody's work on the role of visual tools in social science [1].

The process of generating these histograms leverages random sampling, one of the sampling techniques pivotal to our project's approach to data visualization, ensuring that each visualization provides a representative overview of the dataset, which can be changes as per the user's needs. The efficacy of this sampling method is supported by research from Anderson on "The foundation for understanding the intricacies of random sampling" [3] and Park et al. in their work on "Visualization-aware sampling for very large databases" [12].

Furthermore, the tool's interface, as depicted in the image, includes controls for adjusting the number of bins in the histograms, enabling users to refine their analysis to match specific data exploration needs. This interactivity component is a testament to the tool's design, which adheres to the interactive dynamics for visual analysis taxonomy established by Heer and Shneiderman [11].

In summary, this section will articulate the rich features our tool offers for visual data exploration and the methodological underpinnings that make these features both robust and essential for extracting meaningful insights from complex datasets. The discussion will conclude with a focus on the tool's interactivity and flexibility, hallmarks of modern visual analytics as defined by Keim et al. [17].

4.2.3 User Interaction and Experience

This subsection, encapsulates the interactive dynamics of the visualization tool, spotlighting how users engage with the data through intuitive interfaces and real-time visual feedback mechanisms. This subsection will discuss the user's navigational journey within the tool and the features that enhance the overall data exploration experience.

As the above image provides a clear illustration of the tool's interactive histogram capability. This visualization not only aggregates data into informative bins but also responds to user input, displaying detailed information about each bin's range and frequency upon interaction. This immediate feedback loop is essential for exploratory data analysis, enabling users to quickly identify and focus on areas of the dataset that require further examination [30].

Furthermore, the histogram embodies the tool's commitment to a seamless user experience. The ability to dynamically adjust the number of bins without reloading the page exemplifies the tool's use of progressive disclosure — a design strategy that presents information as needed and reduces cognitive load [2].

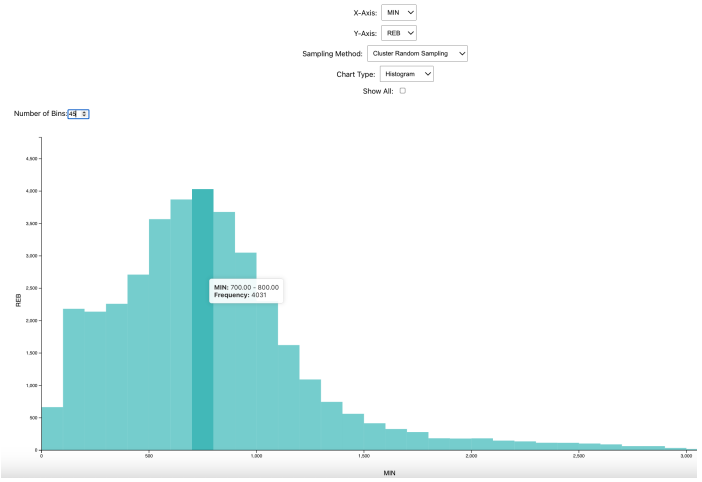

FIGURE 4.3: This histogram provides on-demand data insights, enabling users to interact with the visualization for detailed distribution analysis.

The next image demonstrates the tool's flexibility and user empowerment through the application of various sampling techniques. With a simple interface gesture, users can select a sampling method that best suits their analytical needs, transforming the visualization instantly. This level of interactivity is crucial for comparative analysis, allowing users to discern how different sampling methods can influence the interpretation of data [39].

The interactivity showcased in this image is particularly indicative of the tool's responsive design principles. Users can actively engage with the data, iterating through sampling options and witnessing the immediate effect of these choices on the visual output. This active engagement is a cornerstone of modern visual analytics, supporting a more profound and nuanced understanding of the data [6].

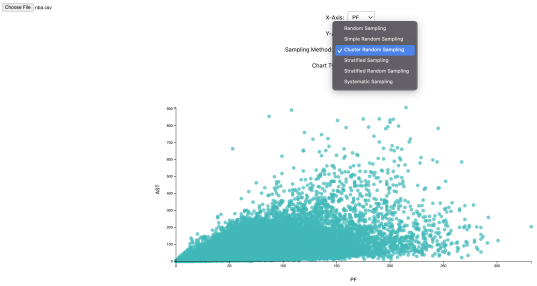

FIGURE 4.4: The tool allows users to choose from various sampling methods, illustrating the impact of different sampling strategies on data visualization.

Together, these images and the accompanying discussion in this subsection will highlight the rich interactive capabilities of the visualization tool. They demonstrate that user interaction is not an afterthought but a primary consideration in the tool's design, reflecting a deep understanding of the visual analysis process and the needs of its users [23].

The subsection will conclude by reiterating the importance of interactivity in the modern data visualization landscape. By enabling users to take an active role in their data exploration, the tool exemplifies the shift from static data presentation to dynamic and interactive data conversation [30].

4.3 D3.js

4.3.1 Overview of D3.js

D3.js, a JavaScript library, is pivotal in the field of data visualization. Its role in this project underscores its versatility in handling complex datasets and creating interactive visualizations. D3.js facilitates a deeper understanding of data by enabling intricate graphical representations.

D3.js stands out for its dynamic data binding capabilities. It allows for intricate visual manipulation of data, making it an ideal choice for projects requiring detailed, interactive visualizations. The library's approach to data-driven transformations and transitions sets a high standard in visual analytics.

In the realm of data analysis, D3.js provides powerful tools for visual exploration. Its ability to handle diverse data types and formats makes it invaluable for projects like this, where data complexity is a major factor. The library's emphasis on web standards ensures compatibility and accessibility [30].

A key aspect of D3.js is its support for interactive visualizations. This interactivity enhances user engagement with data, allowing for a more immersive analytical experience. D3.js's customization capabilities enable the creation of visualizations that are both informative and visually appealing [8].

D3.js's integration into web applications is seamless, allowing for dynamic and responsive visualizations. This project leverages D3.js to create web-based visualizations that are accessible and interactive, showcasing the library's strengths in a web environment [30].

While D3.js offers numerous advantages, its complexity can pose challenges, especially in terms of learning curve and integration with other web technologies. This project addressed these challenges by employing best practices and innovative approaches to maximize the potential of D3.js.

In conclusion, D3.js is a cornerstone of this project's visualization strategy. Its comprehensive feature set, combined with its flexibility and power, makes it an indispensable tool in the data visualization toolkit, particularly for web-based applications.

4.3.2 D3.js in the Project: Implementation, Challenges and Solutions

D3.js played a critical role in this project, primarily used for creating interactive and dynamic visualizations. The implementation involved leveraging D3.js's data-driven approach to bind data directly to the DOM, enabling real-time updates and interactions.

The project capitalized on D3.js's flexibility to create customized visualizations like line charts and scatter plots. These visualizations were tailored to represent the project's complex data sets effectively, demonstrating D3.js's capability to handle diverse data types and formats.

Integrating D3.js with other technologies, especially React, posed significant challenges. The primary issue was managing D3's DOM manipulations within React's virtual DOM. This required a careful balance to ensure both libraries worked together without conflicts.

To address these challenges, a hybrid approach was adopted. React was used to handle the application's architecture and state management, while D3.js focused on the SVG-based rendering and complex visual computations. This separation of concerns allowed for efficient development and performance.

Performance optimization was a key focus, particularly with large data sets. Techniques such as data binding and update optimization in D3.js were employed to ensure that visualizations remained responsive and interactive.

Enhancing user interaction with the visualizations was a crucial aspect. D3.js's event handling capabilities were used to implement interactive features such as tooltips, zooming, and filtering, which enhanced the user experience.

The successful implementation of D3.js in this project highlighted its robustness and versatility as a data visualization tool. The challenges encountered and the solutions developed provide valuable insights into the practical application of D3.js in complex projects.

4.4 React and JavaScript Integration

The integration of React and JavaScript in this project represents a modern approach to building dynamic and responsive web applications. React's component-based architecture, coupled with JavaScript's versatility, played a crucial role in developing interactive and efficient data visualizations. This section will delve into how React and JavaScript were synergized with D3.js, enhancing the project's functionality and user experience. It will explore the technical intricacies of this integration, highlighting the benefits and challenges of combining these powerful technologies to create an engaging data visualization environment.

4.4.1 Combining React with D3.js

In this project, the integration of React and D3.js was a key strategy to create responsive and interactive data visualizations. React's efficient update and rendering mechanisms were used to manage the application's state and UI components, while D3.js handled the detailed visual rendering and data-driven transformations.

The project utilized React to construct the DOM and manage the visualization components, while D3.js was employed for creating and manipulating the SVG elements based on data. This combination ensured smooth transitions and interactions in the visualizations.

One example of this integration can be illustrated with a React component for a D3.js chart. React was used to set up the chart's structure and manage its state, while D3.js handled the SVG creation and data binding.

```
import React, { useEffect, useRef } from 'react';
import * as d3 from 'd3';

function D3Chart() {
    const ref = useRef();

    useEffect(() => {
        const svg = d3.select(ref.current);
        // D3.js chart implementation goes here
    }, []);

    return <svg ref={ref}></svg>;
}
```

This code snippet demonstrates the integration, where a React component uses D3.js within the useEffect hook to manipulate an SVG element.

The main challenge in this integration was handling the overlap between React's and D3's DOM manipulations. The solution involved clearly defining the responsibilities of each library: React for DOM and state management, and D3 for graphical rendering and data-driven transformations.

The integration of React and D3.js in the project represents a powerful combination of technologies for building sophisticated data visualization applications. This approach leverages the strengths of both libraries, offering a scalable and efficient solution for complex data visualization needs.

4.4.2 Project Implementation: Technical Challenges and Solutions

The integration of React and D3.js in the project posed unique challenges, particularly in state management and rendering. React's state management was used to handle data and UI states, while D3.js was responsible for rendering and updating the SVG elements based on this state.

One significant challenge was ensuring synchronization between React's virtual DOM and D3.js's direct DOM manipulation. This required a careful design to ensure that React's rendering lifecycle was not disrupted by D3.js's operations.

Another challenge was maintaining code modularity and reusability. The project adopted a component-based approach, where visualization components were designed as reusable React components integrated with D3.js for the visual rendering part.

Efficient data handling and performance optimization were crucial. The project utilized React's efficient update mechanisms alongside D3.js's data-driven techniques to ensure high performance, even with large datasets.

Enhancing interactivity with complex visualizations brought up challenges in event handling. The solution involved leveraging D3.js for complex event handling while maintaining the overall component structure in React.

Ensuring responsive design and cross-browser compatibility was addressed by combining React's responsive design capabilities with D3.js's SVG-based visualizations, ensuring a seamless user experience across different devices and browsers.

These challenges and their solutions highlight the complexity and intricacies of integrating React with D3.js. The project's approach provides a blueprint for effectively combining these technologies to create sophisticated and interactive data visualizations.

4.5 Visualization Implementations

The Visualization Implementations section of this project focuses on the practical application of the integrated technologies—D3.js, React, and JavaScript—in creating dynamic and interactive data visualizations. This section will explore various visualization types implemented in the project, such as line charts, scatter plots, and histograms. Each type will be examined in terms of its design, functionality, and the specific data it represents. This exploration will not only showcase the technical capabilities of the utilized tools but also demonstrate how they were employed to effectively communicate data insights and enhance user engagement.

4.5.1 Line Chart

Line charts are a fundamental tool in data visualization, used for displaying trends and changes over time [2, 32]. In this project, a line chart was implemented to illustrate temporal data trends, providing a clear and concise view of data changes.

The line chart was implemented using D3.js and integrated within the React framework. D3.js's path generation functions were utilized to plot the data points, while React managed the overall component structure and state, ensuring a responsive and interactive visualization experience.

The implementation involved mapping data points to a line using D3.js's scales and axes functions. React's state management handled the dynamic aspects of the chart, such as data updates and interactivity.

```
import React, { useEffect, useState } from 'react';
import * as d3 from 'd3';

function LineChart({ data }) {
    useEffect(() => {
        const svg = d3.select('#line-chart');
        // D3.js line chart implementation
    }, [data]);

    return <svg id="line-chart"></svg>;
}
```

This code snippet demonstrates how the line chart component was constructed using React and D3.js, showcasing the integration of data binding and SVG manipulation.

One challenge was ensuring the line chart's responsiveness to data changes and window resizing. This was addressed by leveraging React's re-rendering capabilities in response to state changes, combined with D3.js's dynamic SVG updates.

The line chart implementation in this project exemplifies the effective use of D3.js and React to create a dynamic, interactive, and responsive data visualization tool, tailored for time-series data analysis.

4.5.2 Scatter Plot

Scatterplot charts are pivotal in data visualization for examining correlations and patterns between two variables [32]. They plot individual data points on a two-dimensional graph, revealing relationships and distribution trends.

In this project, the scatterplot was implemented using D3.js within a React component. This allowed for dynamic and interactive exploration of data, with D3.js handling the rendering of data points and React managing the chart's state and interactivity.

```
import React, { useEffect } from 'react';
import * as d3 from 'd3';

function ScatterPlot({ data }) {
    useEffect(() => {
        const svg = d3.select('#scatterplot');
        // D3.js scatterplot implementation
    }, [data]);

    return <svg id="scatterplot"></svg>;
}
```

This code snippet highlights the integration of D3.js for SVG-based data point rendering and React for managing the component lifecycle.

The scatterplot required careful mapping of data points to a coordinate system, achieved through D3.js's scale functions. React's stateful components were used to manage dynamic data updates and interactions like hover effects.

A significant challenge was implementing interactive features like tooltips and data point highlighting. This was addressed by combining D3.js's graphical capabilities with React's event handling framework.

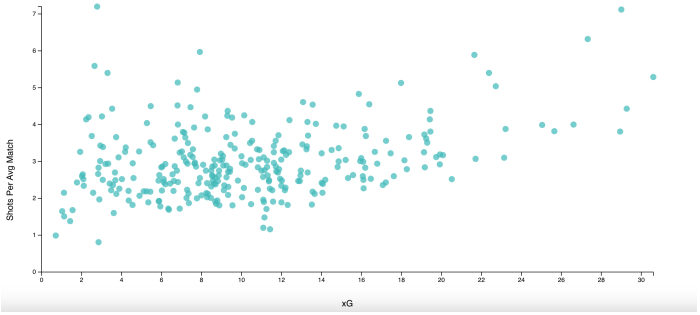

FIGURE 4.5: Scatterplot illustrating the relationship between expected goals (xG) and the average number of shots per match across a dataset. Each point represents aggregated data for an individual team or player, highlighting variations in offensive strategies and efficiencies.

The implementation example sets the stage, but the narrative expands into the broader landscape of data storytelling. The power of a Scatter Plot lies not just in its creation but in its ability to bring hidden patterns to the surface. Visual aids, such as animated graphics illustrating how different datasets manifest as unique scatter patterns, provide a visual journey into the versatility of this technique.

The scatterplot chart in this project demonstrates the effective use of D3.js and React to create an interactive tool for data analysis, showcasing the relationship between different data variables.

4.5.3 Interactive Histogram

Histograms are a type of data visualization used to represent the distribution of a dataset [33]. They provide insights into the frequency of data points within specified ranges, making them essential for understanding the distribution characteristics of the data.

For this project, an interactive histogram was developed using D3.js, integrated within a React framework. This approach enabled the creation of a histogram that not only displayed data distribution but also allowed users to interact with it, such as adjusting bin sizes.

The implementation involved using D3.js to calculate the frequency of data and create the corresponding bars of the histogram. React's state management was used to handle user interactions and dynamically update the histogram.

```
import React, { useEffect } from 'react';
import * as d3 from 'd3';

function Histogram({ data }) {
    useEffect(() => {
        const svg = d3.select('#histogram');
        // D3.js histogram implementation
    }, [data]);

    return <svg id="histogram"></svg>;
}
```

This code snippet illustrates the combination of React for managing the component and D3.js for rendering the histogram.

A challenge in implementing the histogram was ensuring real-time responsiveness to user interactions. React's state management system and D3.js's dynamic rendering capabilities were harnessed to update the histogram efficiently based on user inputs.

The figure represents the distribution of goals scored across matches played from the given dataset. The visualization, created using the application developed in this study, showcases the ability to effectively communicate data trends and patterns in the context of football datasets. The histogram, with a set number of 20 bins, provides a clear depiction of the frequency of goals scored, offering insights into the underlying data through interactive elements integrated via D3.js and React.

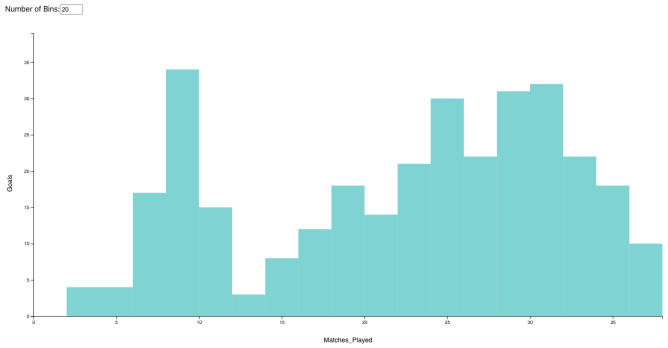

FIGURE 4.6: Histogram representing the distribution of goals scored across matches played with number of bins=20

The interactive histogram developed in this project exemplifies the powerful combination of D3.js and React, enabling the creation of a responsive and user-friendly data visualization tool that provides insights into data distribution.

4.6 Comparative Analysis

This section focuses on a comparative analysis of the visualization tools and techniques used in this project against other available methods in the field of data visualization. It aims to critically assess the advantages and limitations of the chosen tools - D3.js, React, and JavaScript - in the context of the project's specific requirements. This analysis will also explore alternative approaches and technologies, providing a comprehensive overview of how different tools and techniques can impact the outcomes of data visualization projects. The goal is to contextualize the project's choices within the broader spectrum of data visualization practices and innovations.

D3.js's capabilities in creating complex and interactive visualizations are compared with alternatives like Chart.js and Highcharts. While Chart.js offers simplicity and ease of use, D3.js provides more flexibility and control for complex visualizations [30].

React's role in building interactive user interfaces is contrasted with other frameworks like Angular and Vue.js. React's component-based architecture offers distinct advantages in terms of performance and modularity, which is crucial for data visualization applications.

JavaScript's use as a scripting language for web-based applications is compared with other languages like Python and its visualization libraries. JavaScript's integration with web technologies provides a seamless experience for online data visualization projects [11].

The integration strategy of D3.js, React, and JavaScript is evaluated against other combinations like using Vue.js with D3.js or integrating D3.js with Angular, highlighting the unique benefits and challenges of each approach.

This comparative analysis provides insights into the strengths and limitations of the chosen tools and techniques, offering a broader perspective on the various approaches available for data visualization.

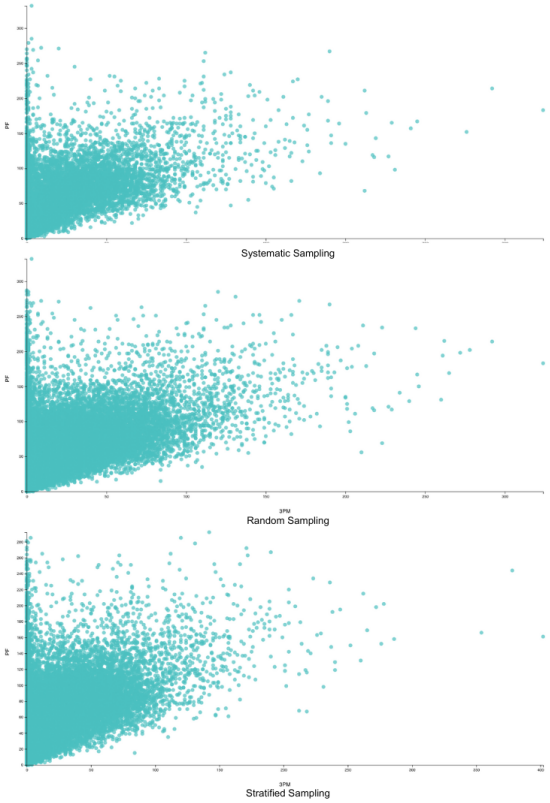

FIGURE 4.7: Three scatter-plot chart using three sampling techniques for the same field of basketball dataset.

The figure illustrates three distinct sampling techniques—Systematic Sampling, Random Sampling, and Stratified Sampling—applied to a dataset. Each subplot presents a scatter plot that visualizes the dispersion of data points according to the method used. In Systematic Sampling, the points display a regular interval pattern, suggesting a fixed sampling sequence. Random Sampling exhibits a more dispersed and unpredictable arrangement of points, indicating the absence of any specific order. Stratified Sampling appears to cluster the data into specific strata or layers, reflecting a deliberate effort to capture representative samples from distinct segments of the dataset. This comparative visualization highlights the differences in data coverage and potential biases introduced by each sampling method.

React's utilization in the project facilitated dynamic user interface management. Its component-based approach allowed for efficient updates and state management, crucial for the interactivity seen in the project's visualizations, such as real-time data updates in the histograms.

JavaScript's role as the foundational programming language was a key advantage. It enabled seamless integration of D3.js and React, creating a unified development environment. This was particularly evident in the project's interactive features, like data filtering and manipulation in the visualizations [11].

The project's integrated approach, combining D3.js for detailed visualizations, React for UI management, and JavaScript for overall functionality, created a synergistic effect. This integration was evident in the fluid interactivity and responsiveness of the visualizations, significantly enhancing user experience.

The combination of these tools was specifically chosen to address the project's needs. For instance, D3.js's data-driven approach was crucial for the project's complex data visualization requirements, while React provided the necessary UI flexibility.

The advantages of using D3.js, React, and JavaScript in the context of this project were evident in the enhanced customization, dynamic UI management, and effective integration, which collectively met the project's specific visualization and interactivity needs.

4.7 Conclusion

The project successfully implemented a series of complex and interactive data visualizations using D3.js, React, and JavaScript. These tools were selected for their unique strengths and integrated to create a cohesive and dynamic user experience.

Throughout the project, various technical challenges were encountered and overcome. This includes integrating D3.js with React's virtual DOM and optimizing performance for handling large datasets, showcasing the flexibility and robustness of these technologies.

A key achievement was the enhancement of user experience through interactive visualizations. This was made possible by leveraging React's efficient UI management alongside D3.js's advanced graphical capabilities.

Looking ahead, the project lays a foundation for further improvements. This could involve exploring newer technologies or refining the current implementation for even better performance and user engagement.

The project's success has broader implications in the field of data visualization. It demonstrates how effectively chosen and integrated tools can transform complex data into insightful and accessible visualizations.

The project's completion was made possible by the collaborative efforts of various individuals and resources. Their contributions were invaluable in navigating the complexities of data visualization.

In conclusion, this project stands as a testament to the power of combining D3.js, React, and JavaScript in creating impactful and interactive data visualizations. The experience and knowledge gained from this project are significant contributions to the field of data visualization.

Chapter 5

Conclusions and Future Work

5.1 Summary: Main Objectives and Findings

This chapter encapsulates the entirety of the study, revisiting the main objectives, findings, and the contributions of the project. It serves to provide a consolidated overview of the research, highlighting key insights and the value added to the field of data visualization. The summary will succinctly recapitulate the project's journey, from its inception to its conclusion, underscoring the significant milestones and the knowledge gained throughout the process. This reflection sets the stage for a detailed discussion on the strengths, weaknesses, future potential, and impact of the study.

The primary objective of this study was to provide a summary view of sampled data from large datasets. This involved utilizing advanced data visualization techniques to make complex data more accessible and interpretable.

Using D3.js, React, and JavaScript, the project effectively created interactive visualizations like scatter plots and histograms. These tools enabled a detailed yet comprehensible representation of large datasets, allowing for efficient data analysis and insight generation.

A key aspect of the project was the implementation of various data sampling techniques. These techniques were crucial in managing the size and complexity of the datasets, ensuring that the visualizations remained efficient and informative. The integration of interactive features into the visualizations significantly enhanced the analytical capabilities of the project. Users could interact with the data in real-time, gaining deeper insights through dynamic visual exploration.

While specific code implementations are extensive, they broadly encompass the integration of D3.js for creating visualizations, React for managing user interface components, and

JavaScript for handling data processing and application logic. The findings from this project demonstrate the effectiveness of combining multiple technologies for data visualization. The study shows how well-designed visualizations can aid in making sense of large and complex datasets.

This project contributes to the field of data visualization by showcasing practical applications of modern tools and techniques in handling big data. It provides a framework for future projects that require efficient and effective data analysis.

The study successfully met its objectives, providing valuable insights into the potential of data visualization tools in analyzing large datasets. The findings underscore the importance of integrating various technologies for effective data analysis.

The project's integration of D3.js, React, and JavaScript represents a significant contribution to the field of data visualization, especially in handling large datasets. This integration illustrates a practical approach to creating dynamic and interactive visualizations.

Implementing advanced data sampling techniques for effective data visualization is a notable contribution. These techniques allow for efficient analysis of large datasets, ensuring the visualizations are both manageable and informative. The development of interactive visualizations such as scatter plots and histograms using these technologies has enhanced the user's analytical experience, making complex data more accessible.

The project's approach to visualization offers a high degree of customization and flexibility. This is particularly evident in the tailored solutions provided for different types of data representations. Addressing the challenge of performance optimization in visualizing large datasets is another key contribution. The project demonstrates effective strategies for managing and rendering large amounts of data without compromising on performance.

Enhancing the user experience through interactive and responsive design has been a focus of the project. This has been achieved by leveraging the strengths of the integrated technologies. The project provides a comprehensive framework that can be adapted for future data visualization projects. It serves as a guide for effectively using modern technologies in data visualization.

These contributions showcase the project's success in pushing the boundaries of data visualization, particularly in the context of large and complex datasets. They represent a significant advancement in the field, offering valuable insights and methodologies for future research and development.

5.2 Critical Analysis

This section of the applied research project delves into a critical analysis of the project, evaluating its strengths and weaknesses. It aims to provide an objective assessment of the methodologies employed, the technologies used, and the overall execution of the project. This analysis will help in identifying areas where the project excelled and where there were shortcomings or challenges. Understanding these aspects is crucial for gaining a comprehensive view of the project's impact and for guiding future work in the field of data visualization.

The project's integration of D3.js, React, and JavaScript is its most significant strength. This combination allowed for the creation of highly interactive and customizable data visualizations, facilitating deep insights into large datasets. The use of React enhanced the user interface's responsiveness and interactivity, making the visualizations more engaging and accessible [19].

A notable weakness lies in the complexity of integrating D3.js with React. This complexity presents a steep learning curve, potentially limiting the project's accessibility to developers with specific skill sets. Additionally, managing the interplay between D3.js's DOM manipulations and React's virtual DOM posed significant challenges. While effective for large datasets, the project faced limitations when dealing with extremely large or complex datasets. Performance issues such as longer loading times and decreased responsiveness were observed, indicating a need for further optimization [5].

Another challenge was balancing customization with usability. While D3.js offers extensive customization, it can sometimes come at the cost of user-friendliness and ease of implementation. The project's reliance on data sampling techniques, though effective in managing large datasets, raises concerns about data accuracy and representation. Ensuring that the sampled data accurately reflects the larger dataset is crucial for the validity of the visualizations [40].

The complex nature of the visualizations, while insightful, could potentially overwhelm users not familiar with advanced data analysis, suggesting a need for more intuitive navigation and data representation.

In summary, the project successfully demonstrates the effective use of modern data visualization tools, though it encounters challenges in complexity, performance optimization, and user experience. These insights provide valuable lessons for future projects in similar domains.

5.3 Future Work

Future research can explore integrating advanced machine learning techniques to enhance data sampling and visualization. This could involve using AI algorithms for predictive analysis and pattern recognition in large datasets. Another area of development is the incorporation of real-time data processing, which would significantly improve the application's responsiveness and usability [41].

Methodological improvements should focus on optimizing performance for extremely large datasets. This could involve refining data processing and rendering techniques or exploring more efficient algorithms for data sampling [5]. Enhancing user experience is another key area, possibly by developing more intuitive interfaces and navigation tools to make the visualizations accessible to a wider audience.

Expanding the project's compatibility across various platforms and devices is crucial. Efforts could be directed towards making the visualizations more responsive and adaptable to different screen sizes and resolutions [11]. Incorporating user feedback mechanisms to gather insights on usability and functionality can guide future improvements. This feedback can be instrumental in refining the visualizations and overall user experience.

Exploring a wider range of visualization types and styles can broaden the project's applicability. This might include geographic data representations or more complex 3D visualizations [30].

Future work in this project holds immense potential for growth and innovation. By focusing on these areas, the project can evolve to meet emerging challenges and demands in the field of data visualization.

5.4 Final Remarks

Reflecting on the project's journey, it stands as a testament to the power of integrating advanced technologies in data visualization. The successful combination of D3.js, React, and JavaScript has paved the way for new possibilities in the realm of big data analysis.

The project's success was not just a result of technological innovation but also a collaborative effort. It underscores the importance of teamwork and shared expertise in tackling complex challenges in data visualization.

This project contributes significantly to the field of data visualization, especially in handling and interpreting large datasets. It serves as a valuable reference point for future projects and research in similar domains [13]. The learning curve experienced throughout the project

was steep yet rewarding. It highlights the importance of continuous learning and adaptation in the ever-evolving field of technology.

Looking forward, the project lays a foundation for further innovation and exploration in data visualization. It opens doors to new research, development opportunities, and methodological improvements, promising continued advancement in the field [41].

In conclusion, this project is a milestone in data visualization, combining technological expertise with a deep understanding of data analysis. It stands as a significant contribution to the field, inspiring future work and ongoing exploration.

Bibliography

[1] Kieran Healy and James Moody. Data visualization in sociology. *Annual Review of Sociology*, 40(1):105–128, 2014. doi: 10.1146/annurev-soc-071312-145551. URL https://doi.org/10.1146/annurev-soc-071312-145551.

[2] E.R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 2001. ISBN 9781930824133.

[3] T. Anderson. The foundation for understanding the intricacies of random sampling, providing crucial insights for our exploration of sampling techniques. In *Introduction to Random Sampling.*

[4] William G. Cochran. *Sampling Techniques, 3rd Edition.* John Wiley, 1977. ISBN 0-471-16240-X.

[5] Mohammad Sultan Mahmud, Joshua Zhexue Huang, Salman Salloum, Tamer Z. Emara, and Kuanishbay Sadatdiynov. A survey of data partitioning and sampling methods to support big data analysis. *Big Data Mining and Analytics*, 3(2):85–101, 2020. doi: 10.26599/BDMA.2019.9020015.

[6] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. 03 2008. doi: 10.1007/978-3-540-70956-5_7.

[7] Sharon L Lohr. *Sampling: design and analysis.* CRC press, 2021.

[8] Scott Murray. *Interactive data visualization for the web: an introduction to designing with D3.* " O'Reilly Media, Inc.", 2017.

[9] Kieran Healy. *Data visualization: a practical introduction.* Princeton University Press, 2018.

[10] Philip Sedgwick. Stratified cluster sampling. *Bmj*, 347, 2013.

[11] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue*, 10(2): 30–55, 2012.

[12] Yongjoo Park, Michael Cafarella, and Barzan Mozafari. Visualization-aware sampling for very large databases. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 755–766, 2016. doi: 10.1109/ICDE.2016.7498287.

[13] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile networks and applications*, 19:171–209, 2014.

[14] Scott Murray. *Interactive data visualization for the web: an introduction to designing with D3*. " O'Reilly Media, Inc.", 2017.

[15] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013. doi: 10.1109/TVCG.2013.126.

[16] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. Effectiveness of animation in trend visualization. *IEEE transactions on visualization and computer graphics*, 14(6):1325–1332, 2008.

[17] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[18] Thomas M Annesley. Bars and pies make better desserts than figures. *Clinical chemistry*, 56(9):1394–1400, 2010.

[19] Lam Thanh Tung Vo. Web application development with react and google firebase: data visualization. 2020.

[20] Susan Gardner Archambault, Joanne Helouvry, Bonnie Strohl, and Ginger Williams. Data visualization as a communication tool. *Library Hi Tech News*, 32(2):1–9, 2015.

[21] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue*, 10 (2):30–55, feb 2012. ISSN 1542-7730. doi: 10.1145/2133416.2146416. URL https://doi.org/10.1145/2133416.2146416.

[22] Howard Brody, Michael Russell Rip, Peter Vinten-Johansen, Nigel Paneth, and Stephen Rachman. Map-making and myth-making in broad street: the london cholera epidemic, 1854. *The Lancet*, 356(9223):64–68, 2000.

[23] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[24] William Playfair. *Playfair's commercial and political atlas and statistical breviary*. Cambridge University Press, 2005.

[25] Ian Spence. William playfair and the psychology of graphs. In *Proceedings of the American Statistical Association, Section on Statistical Graphics*, pages 2426–2436, 2006.

[26] Julie Steele and Noah Iliinsky. *Beautiful visualization: Looking at data through the eyes of experts.* " O'Reilly Media, Inc.", 2010.

[27] Richard Arnold Johnson, Dean W Wichern, et al. Applied multivariate statistical analysis. 2002.

[28] Tamara Munzner. *Visualization analysis and design.* CRC press, 2014.

[29] Robert Lang. Using gapminder. *GW-Unterricht*, 126:76–87, 2012.

[30] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[31] Interactive business dashboard example. https://www.datapine.com. Accessed: 2023-12-08.

[32] William S Cleveland. *The elements of graphing data.* Wadsworth Publ. Co., 1985.

[33] William S Cleveland. *Visualizing data.* Hobart press, 1993.

[34] Alberto Cairo. *The Functional Art: An introduction to information graphics and visualization.* New Riders, 2012.

[35] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185.

[36] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[37] Fan Bao and Jia Chen. Visual framework for big data in d3.js. In *2014 IEEE Workshop on Electronics, Computer and Applications*, pages 47–50, 2014. doi: 10.1109/IWECA.2014.6845553.

[38] NASA. Eyes on Asteroids, 2023. URL https://eyes.nasa.gov/apps/asteroids/#/home. [Online; accessed 10-May-2023].

[39] Stephen Few and Perceptual Edge. Data visualization: past, present, and future. *IBM Cognos Innovation Center*, pages 1–12, 2007.

[40] Ajay S Singh and Micah B Masuku. Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of economics, commerce and management*, 2(11):1–22, 2014.

[41] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. Making data visualization more efficient and effective: a survey. *The VLDB Journal*, 29:93–117, 2020.