R Markdown Trial

Selin Aytan , Idil Bukre Tuzkaya

2025-07-28

Contents

PREFACE

Motivation and Goal

This booklet was born out of a growing need to communicate complex data in a clear and accessible way. As data grows in volume and complexity, visual tools have become essential for uncovering insights that might otherwise remain hidden.

Motivated by a personal interest in data visualization, this project also aims to fill the gaps left by existing resources—many of which focus either too heavily on code or abstract theory. Our goal is to provide a guide that is technically sound yet simple and intuitive, aimed at readers with a basic understanding of data.

While various tools exist for visualization, this booklet focuses primarily on R, emphasizing its strengths in flexibility, reproducibility, and statistical integration. We also briefly compare R to other tools where relevant, to highlight why it was chosen and how it supports effective data storytelling from preparation to final presentation.

Challenges and Solutions During the Project

Throughout the project, we encountered several challenges—both technical and conceptual. One of the first difficulties was handling large and complex datasets efficiently within the R environment. Some of the packages required for big data visualization demanded careful memory management and unfamiliar workflows, especially when working with tools like SparkR and Arrow.

Another major challenge was choosing the right type of visualization for different kinds of data and ensuring that the graphics remained both accurate and interpretable. It often took multiple iterations to find the right balance between simplicity and completeness.

We also spent considerable time understanding the underlying structure of the data and preparing it for analysis. Missing values, inconsistent formats, and merging different data sources required a lot of manual adjustment and trial-and-error.

Despite these challenges, we approached each problem methodically—consulting documentation, testing alternative solutions, and learning from examples and community discussions. These experiences not only strengthened our understanding of data visualization in R, but also improved our overall problem-solving skills in working with real-world data.

Target Audience

This booklet is intended for students, early-stage researchers, and data enthusiasts who want to strengthen their understanding of data visualization using R. It assumes only a basic familiarity with data analysis and aims to present both core principles and practical examples in a clear and accessible way, making it useful for both beginners and those looking to refresh their skills.

Thanks

This project was carried out as part of a summer internship at the University of Edinburgh, supported by the Erasmus+ programme. We would like to thank Ozan Evkaya, PhD for his valuable guidance and support throughout the process. His feedback and expertise greatly contributed to the development of this booklet.

1.Introduction

1.1 Purpose and Objectives of the Study

In today's world, with the explosive growth of data, deriving valuable insights and presenting them in a meaningful way has become ever more essential. Particularly, when working with big data, statistical analysis alone is often insufficient; supporting these analyses with powerful visualizations should be regarded as key priority in data-driven decision making.

This project aims to explore how classic visualization techniques are applied to "small" to "moderate" datasets to the realm of big data. Our project aims to demonstrate that data visualization is not merely about creating graphs, but about generating meaningful insights. It explores which types of graphs are suitable for different data structures, how to construct them using small to moderate datasets, and how these techniques applied when it comes to big data. Furthermore, the interactive dashboards and visualizations we used in this booklet targets foster engagement with reader. By doing so, we aim to contribute open-source materials and support R community. Main focuses of the project are:

- Improving practical skills while dealing with big data.
- Utilizing R tools such as Spark, Arrow, ggplot2, and tmap in order to create scalable and effective visual representations.
- Highlighting the differences between different types of data, and appyling proper visualizations for each.
- Safeguarding that visualizations are accessible and interpretable for both technical and general users.
- Promoting scientific transparency by designing reproducible analysis process.

Through this project, we intend to bridge the gap between big data and data visualization, illustrating how technical tools in data science can be used for powerful representations.

1.2 Overview of Datasets Used

In this Project ... datasets are used. Here is the overview:

Scottish Index of Multiple Deprivation 2020: The Scottish Index of Multiple Deprivation is a relative measure of deprivation across 6,976 small areas (called data zones). If an area is identified as 'deprived', this can relate to people having a low income but it can also mean fewer resources or opportunities. SIMD looks at the extent to which an area is deprived across seven domains: income, employment, education, health, access to services, crime and housing. The dataset description was obtained from the Scottish Government's official SIMD documentation[1].

[1]https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/

1.3 Project Roadmap

This project follows a systematic approach: starting from data comprehension and preparation, moving on to creating effective visualizations, and finally scaling them to a big data context while ensuring reproducibility. Below is a an overview of the major stages:

1. Data Exploration, Understanding, and Preprocessing

In this step, EDA(Exploratory Data Analysis) was performed to gain a solid understanding of the datasets' fundamental characteristics. This process involved handling missing values, assessing normality, and identifying data types. In order to apply various visualization techniques, we deliberately focused on datasets containing diverse data types. Main focus was primarily on Scotland based data, such as deprivation and Fringe datasets. In the end, the data was cleaned and prepared for visualization.

2.Initial Visualizations with Small/Moderate Datasets

For this step, we followed an iterative approach to decide which type of visualizations were suitable for each dataset in terms of aesthetics and readability. Testing these techniques initially on smaller datasets provided remarkable insights into how appropriate they scale in context of big data.

3.Scaling Visualizations to Big Data

4.Spatial and Interactive Visualizations

5.Documentation and Reproducibility

6.Final Outputs and Contribution

2.R Environment and Package Ecosystem

2.1 Why R for Data Visualization?

R is considered to be one of the most powerful tools and has a gained significant audience for a variety of factors ranging from its flexibility to rich package ecosystem to built-in integration of statistical computing. ıt'S features also include very efficient data handling and storage facilities. R boasts a comprehensive ecosystem of packages designed spesifically for data visualization.R is also open sourced and compasses a large and active community of both users and developers. r also integrates quite well with other programming languages like Python and SQL. it also aloows for embedding of R outputs in web applications through frameworks like Shiny, making it easier to share your visualizations in interactive web dashboards or other software applications.In academic and professional settings, the ability to reproduce analyses is crucial. R scripts can be shared and rerun, providing an auditable trail of how your visualizations were created. This is particularly important for transparency in research and reporting.and lastly being an open-source program platform ,R is run without any cost.This makes it a cost-effective solution for data visualization, as there are no licensing fees involved, making it accessible for students, researchers, and businesses alike