VISUALISING HISTORICAL GENDER DIVERSITY IN COMPUTER SCIENCE PUBLICATIONS

By

Vamsi Mudila
(201668163)

A DISSERTATION

Submitted to

The University of Liverpool

in partial fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

Primary Supervisor – Patrick Totzke

Secondary Supervisor – Meng Fang

SEPTEMBER 2023

ABSTRACT

VISUALISING HISTORICAL GENDER DIVERSITY IN COMPUTER SCIENCE PUBLICATIONS

By

Vamsi Mudila

This dissertation explores the project "Visualizing Historical Gender Diversity in Computer Science Publications." The project's core aim was to analyse and visualize gender diversity trends in computer science by examining authorship data from the DBLP API. This abstract summarizes the project's objectives, methods, key findings, and contributions concisely. The project's primary goal was to investigate gender representation in computer science research publications over time. To achieve this, the data was collected and cleaned from the DBLP API, focusing on author names, publication venues, and publication years. To accurately assign genders to authors, a gender classification approach was developed, considering diverse gender identities. This classification was vital for meaningful analysis of gender diversity. Statistical methods were employed to analyse the data, revealing trends in gender representation across time and subfields of computer science. Results indicated progress in recent years but also persistent gender disparities in certain areas. A notable outcome of the project is an interactive web application that enables users to explore gender diversity trends easily. The application offers user-friendly visualizations like donut charts, line plots, and drop-down menus, enhancing data interpretation. In summary, this dissertation sheds light on historical gender diversity in computer science. It emphasizes the importance of addressing gender disparities while providing a practical tool for researchers, educators, and policymakers to further investigate and promote diversity and inclusion in computer science.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of another.

I confirm that I have not copied material from another source nor committed plagiarism nor commissioned all or part of the work (including unacceptable proof-reading) nor fabricated, falsified, or embellished data when completing the attached piece of work.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Vamsi Mudila

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to the following individuals and entities who have played a significant role in the completion of this dissertation.

I am deeply thankful to the University of Liverpool for providing me with a conducive learning environment and invaluable resources throughout my academic journey. This institution has been the foundation of my knowledge and growth.

I extend my sincere appreciation to my professors for their unwavering guidance, expert insights, and continuous support. Their mentorship has been instrumental in shaping my academic pursuits.

I would like to acknowledge the exceptional guidance provided by Dr. Patrick Totzke and Dr. Meng Fang whose expertise and dedication have been pivotal in the successful completion of this dissertation. Their constructive feedback and encouragement have been invaluable.

To my friends, who have stood by me with unwavering support and understanding, I express my deepest gratitude. Your camaraderie and encouragement provided a sense of balance and motivation during challenging times.

Lastly, I owe an immeasurable debt of gratitude to my family. Their boundless love, encouragement, and belief in my abilities have been my greatest source of strength. This dissertation would not have been possible without their unwavering support.

I would also like to acknowledge the contributions of all those who, in various ways, have assisted me during this academic endeavour. Your support, no matter how minor, is greatly appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1. INTRODUCTION

In our ever-changing world driven by rapid technological progress, computer science plays a crucial role in shaping our future. As this field continues to expand, it's essential to explore the dynamics of gender diversity within it. This introduction sets the stage for understanding the project's goals, methods, and its significant impact on knowledge.

1.1. Project Description

1.1.1. Non-Technical Summary:

This project aims to dive deep into the world of computer science research, unravelling the complex tapestry of gender diversity throughout its history. At its core, this endeavour seeks to answer fundamental questions: Who are the authors behind the research papers that influence computer science, and how has the gender makeup of these authors changed over time? Additionally, are there noticeable differences in gender representation in various subfields of computer science? To make these insights accessible to everyone, a user-friendly website to visualize the data and make exploration easy has been created.

1.1.2. Technical Summary:

By leveraging the DBLP API, a publicly available repository filled with computer science research papers, this project embarks on a multifaceted journey. Initially, I meticulously extract essential paper details, including author information, publication venues, and timestamps. However, I face a significant challenge: accurately determining authors' genders based solely on their names. Addressing this intricate task requires advanced natural language processing and data analysis techniques. Once I identify genders, robust statistical methods come into play, allowing us to uncover patterns and trends in gender representation throughout the history of computer science. The result is a collection of visually intuitive graphs and charts presented on a user-friendly website.

1.2. Contribution to Knowledge

This project stands as a substantial contribution to the ongoing conversation about gender diversity in the vast field of computer science. By meticulously analysing the gender composition of research paper authors, it aims to spark discussions and inspire actions aimed at promoting gender equity in this field.

The resulting website serves as an invaluable tool for anyone interested in gaining profound insights into the complex landscape of gender diversity within computer science research publications.

1.3. Solution and Effectiveness

The solution produced by this project, in the form of a user-friendly website designed to visualize gender diversity trends in computer science research, represents a significant contribution to the field. This tool empowers users to explore data-driven insights, identify historical shifts, and gain a deeper understanding of gender representation within the domain of computer science.

The effectiveness of this solution is multi-faceted. It can be measured not only by its accessibility and ease of use but also by its potential to stimulate meaningful discussions and inspire actions aimed at promoting gender diversity and inclusivity in computer science. By shedding light on historical and current patterns, the website equips stakeholders with the knowledge needed to make informed decisions and advocate for equitable representation. Furthermore, its impact extends beyond academia, serving as a valuable resource for anyone interested in understanding and addressing gender disparities in technology-related fields.

Chapter 2. AIMS AND OBJECTIVES

Aims:

1. Analyse how gender representation in computer science research has evolved over time by examining authorship. (Achieved)

2. Build an interactive website to visually display this information, enabling easy exploration and understanding. (Achieved)

Objectives:

1. Gather and clean data from the DBLP API database for analysis. (Achieved)

2. Identify a method to accurately assign gender to authors, while respecting all genders including non-binary and non-traditional identities. (Achieved)

3. Apply statistical methods to track trends in gender representation over time and across computer science fields, verifying findings with statistical tests. (Achieved)

4. Create comprehensive graphs and charts that effectively communicate the findings to a broad audience. (Achieved)

5. Develop an easy-to-use website incorporating these visuals, allowing users to filter and explore data as per their interests. (Achieved)

6. Adhere to ethical guidelines and data privacy regulations throughout, anonymizing data where necessary and respecting all gender identities. (Achieved)

Chapter 3. BACKGROUND

3. 1. Literature Review

Gender diversity in computer science (CS) has gained attention in recent years, with a focus on understanding gender representation in CS publications. This review summarizes key research studies and their findings, methodologies, and potential conflicts in viewpoints.

Key Findings in Gender Diversity Research:

- **Michael Ley (2009): "DBLP: Some Lessons Learned"** [1] **:**

  This study emphasizes the importance of the DBLP database as a source for CS publications. While it doesn't directly address gender diversity, it highlights the evolving landscape of CS research contributors.

- **Karimi, F. et al. (2016): "Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods"** [2] **:**

  This research explores methods to predict gender from web names, showcasing the potential of automated techniques for gender inference.

- **Vasilescu, B. et al. (2015): "Gender and Tenure Diversity in GitHub Teams"** [3] **:**

  Vasilescu and their teams work investigates gender and tenure diversity in GitHub teams, revealing disparities in contributions and acceptance rates. It combines data analysis and surveys.

- **Correll, S. J. (2001): "Gender and the Career Choice Process: The Role of Biased Self-Assessments"** [4] **:**

  Correll's research examines the influence of biased self-assessments on gender-related career choices, especially in CS-related fields, using surveys and interviews.

- **Wang, M.-T. and Degol, J.L. (2016): "Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions"** [5] **:**

  This literature review provides an overview of the gender gap in STEM fields, offering insights into the current state of knowledge and implications for practice.

- **West, J.D., et al. (2013): "The Role of Gender in Scholarly Authorship"** [6] **:**

  West and colleagues examine gender disparities in scholarly authorship, revealing a persistent gap in academic publishing through data analysis.

- **Holman, L., Stuart-Fox, D., Hauser, C.E. (2018): "The Gender Gap in Science: How Long Until Women Are Equally Represented?"** [7] **:**

  This study discusses the gender gap in science and predicts future gender parity through mathematical modelling and data analysis.

3.2.2. Methodological Approaches and Conflicting Viewpoints:

The research presented above employs various methodologies, including data analysis, machine learning, surveys, and mathematical modelling, to examine gender diversity in CS publications. While these studies contribute valuable insights, there are several conflicting viewpoints and unresolved issues in the literature:

1. **Data Quality and Representativeness:**

   Some studies rely on data sources like DBLP, which may not accurately capture gender diversity. This limitation can introduce potential biases in research findings, highlighting the need for improved data sources.

2. **Intersectionality:**

   Many studies primarily focus on gender but may not consider intersectional factors such as race, ethnicity, or socioeconomic background. Addressing these intersectional dimensions of diversity is critical for a more comprehensive understanding of the issue.

3. **Ethical Considerations:**

   The ethical implications of inferring gender from names or using gender-related data in research are subjects of ongoing debate. Ensuring the responsible and ethical use of gender-related data is a significant concern in this research domain.

4. **Cultural and Regional Variations:**

   Research in this area may not always account for cultural or regional variations in gender dynamics within computer science. These variations can influence the extent and nature of gender diversity challenges.

In conclusion, this literature review demonstrates that gender diversity in CS publications is a complex and multifaceted issue. While existing research provides valuable insights, there are still significant gaps and challenges to address.

The next section discusses how this project aims to fill some of these gaps and contribute to a more nuanced understanding of gender diversity in CS publications.

3. 3. Research Gaps

In the existing research about gender diversity in computer science (CS) publications, some gaps and limitations need addressing. Here's what these gaps are and how the project, intends to bridge them, highlighting the uniqueness and importance of its approach:

1. **Limited Gender Attribution in CS Publications Data:**

   Research Gap: Many previous studies face a big challenge because CS publication databases like DBLP don't explicitly mention authors' genders.

   Project Significance: This project stands out by developing a special method to attribute genders based on authors' names. It aims to enhance gender-related analysis in CS publications. This innovative approach can greatly improve the accuracy and depth of gender diversity research.

2. **Data Visualization for Enhanced Understanding:**

   Research Gap: While some existing studies provide valuable insights, the presentation and visualization of gender diversity data can be further improved.

   Project Significance: This project incorporates advanced data visualization techniques, drawing from principles in data science and visualization. It strives to present findings in a more accessible and informative manner, making complex data more understandable to a broader audience. This aspect is pivotal in facilitating informed decision-making, policy development, and awareness about gender diversity issues in CS.

In summary, this project aims to address critical research gaps related to gender attribution in CS publications data and enhance data visualization techniques.

By filling these gaps, the project has the potential to foster a more inclusive and equitable research environment in computer science, ultimately benefiting academia and society.

3. 4. Relevant Theories and Concepts

To understand gender diversity, representation, and inclusion in STEM fields, especially computer science, it's important to introduce and explain some key theories and concepts:

1. **Social Identity Theory** [8] **:**

   Explanation: This theory suggests that individuals categorize themselves into social groups based on characteristics like gender, deriving a sense of identity and self-esteem from these groups.

   Relevance: It helps us understand how individuals identify with their gender group and how this identification can affect their experiences in STEM fields.

2. **Stereotype Threat** [9] **:**

   Explanation: Stereotype Threat occurs when individuals, aware of negative stereotypes about their group (e.g., women in STEM), may underperform due to fear of confirming those stereotypes.

   Relevance: It's crucial for understanding the barriers women face in computer science and other STEM disciplines.

3. **Inclusive Excellence** [10] **:**

   Explanation: This framework emphasizes that diversity and inclusion are essential for excellence in higher education and research.

   Relevance: It underscores the importance of an inclusive environment in STEM where individuals from diverse backgrounds, including women, can contribute.

4. **Implicit Bias** [11] **:**

   Explanation: Implicit Bias refers to subconscious attitudes or stereotypes that influence people's decisions and actions, often without their awareness.

   Relevance: Recognizing and addressing implicit bias is critical for promoting gender diversity and equity in STEM.

5. **Pipeline Problem** [12] **:**

   Explanation: This theory suggests that underrepresentation of certain groups, like women, in STEM fields results from disparities at various education and career stages.

Relevance: Understanding the Pipeline Problem is essential for developing targeted interventions to increase the participation of women in computer science.

6. **Growth Mindset** [13] **:**

   Explanation: The Growth Mindset concept posits that individuals who believe they can develop their abilities through effort and learning are more successful and resilient.

   Relevance: Promoting a Growth Mindset can create a more inclusive environment in STEM, encouraging women to persist despite challenges.

These theories and concepts provide a theoretical framework for understanding gender diversity challenges and opportunities in computer science and STEM fields, enriching the context of your research.

Chapter 4. Ethical Use of Data

4.1. Data Sources

For this research project, two primary data sources were utilized to examine historical gender diversity in computer science publications:

1. **DBLP Database** [14] **:** The DBLP database is a comprehensive resource that contains bibliographic information about computer science publications. It includes data on authors, publications, and their affiliations. Access to this data is openly available to the public.

2. **Kaggle Dataset** [15] **:** The Kaggle dataset used for gender prediction is derived from publicly available information, specifically names. This dataset is anonymized and does not contain any personally identifiable information.

4.2. Ethical Considerations

The ethical use of data is of utmost importance in this project. It is crucial to note that both data sources used in this research project meet ethical standards:

1. **DBLP Database:** The DBLP database is publicly accessible and does not contain sensitive or personal information about individuals. It is openly available for academic and research purposes, making its use in this project ethically sound.

2. **Kaggle Dataset:** The dataset I used to predict gender is derived from publicly available names. It's anonymized, meaning it doesn't reveal anyone's identity, and it doesn't contain private or confidential information. Therefore, it's ethically sourced.

4.3. Ethical Approval

As both data sources used in this project are openly available and do not contain sensitive personal data, there was no need to obtain specific ethical approval for data collection. However, it is important to mention that ethical considerations were considered throughout the project to ensure the responsible and ethical use of data.

It's essential to emphasize that no personal or sensitive information was gathered or used in this research. The project's focus was on historical gender diversity trends in computer science publications, and all data used was in the public domain or anonymized.

Chapter 5. DESIGN

5.1. Design Overview

The design of the research project involves several key components and stages that collectively form a well-structured and comprehensive system. This section provides an in-depth view of the project's design, including its architecture, data flow, and user interface.

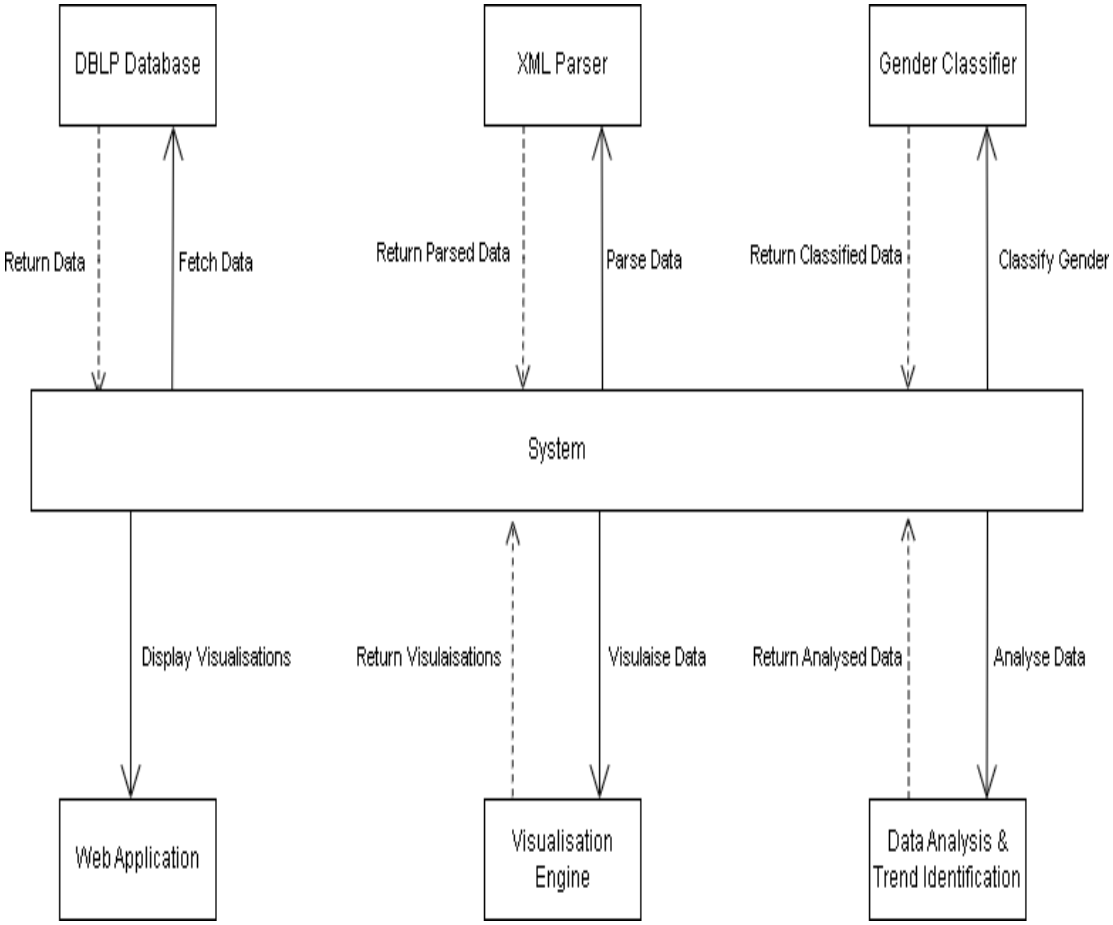
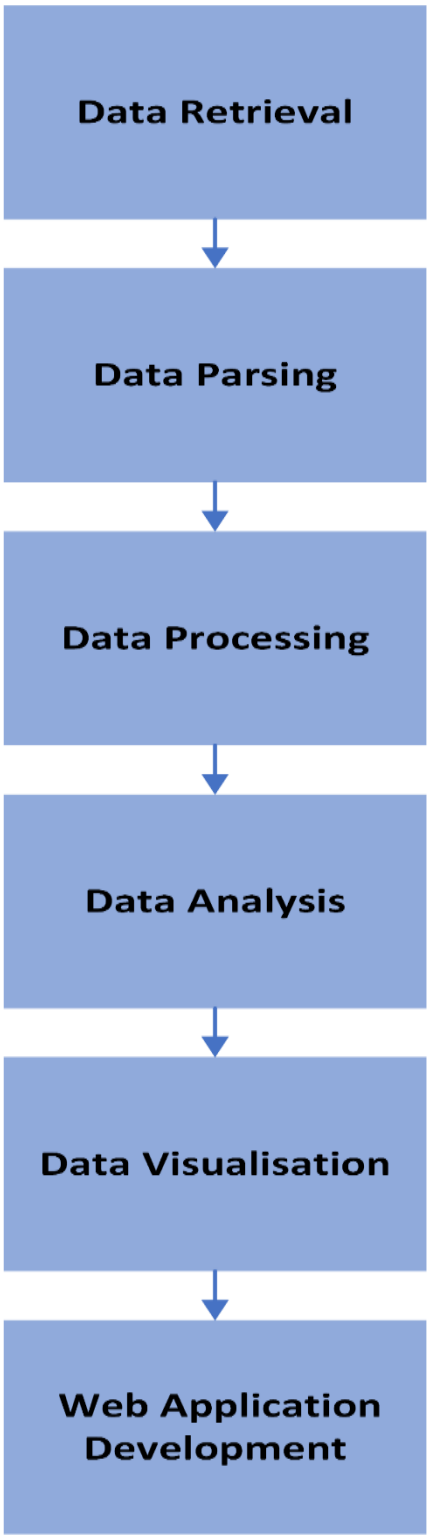


Figure 1 - In-depth UML Diagram

5.2. System Architecture



Connection to the DBLP database is established to fetch the required raw data.

The raw data is parsed into a structured format using a custom XML parser.

Author names are classified by gender using either a Gender API or a custom machine learning model.

Trends in gender representation in computer science research are identified through data analysis using Python's Pandas, NumPy, and SciPy libraries.

Visualizations of the trends are created using Matplotlib and Plotly to represent the analysed data.

A user-friendly web application is developed to enable users to interact with the visualized data and findings.

Figure 2 - Block Diagram of a high-level overview design

5.3. Pseudocode for Key Algorithm

The main logic of the system involves processing and analysing data. A simplified pseudocode for this procedure might look like this:

```
Algorithm Gender_Classification
Input: Parsed data with author names
Output: Data with assigned genders

1  function Gender_Classification(parsed_data)
2      for each entry in parsed_data do
3          name <- extract author name from entry
4          gender <- call GenderAPI(name) or predict using ML model
5          add gender to entry
6      end for
7      return parsed_data with gender
8  end function
```

This pseudocode represents the process of iterating over each entry in the parsed data, extracting the author's name, determining the gender (either through an API call or a prediction from a machine learning model), and then adding that gender information back to the entry. The result is the same dataset, but now with an additional piece of information: the gender of each author.

5.4. User Interface Mockup

1. Navigation Bar: It is at the top. It helps users go to various parts of the site like Home, About, and Contact.

2. Home Page: It is the first page users see. It quickly tells them what the website is for and guides them where to go next.

3. Content Pages: These pages hold the main information or features of the website. For example, they may show graphs or tables.

4. About Page: This page tells users about the purpose of the web application and who created it.

5. Contact Page: Users can find how to reach the creator of the website here.

6. Footer: It is at the bottom. It may include links to privacy policies, terms and conditions, and social media.

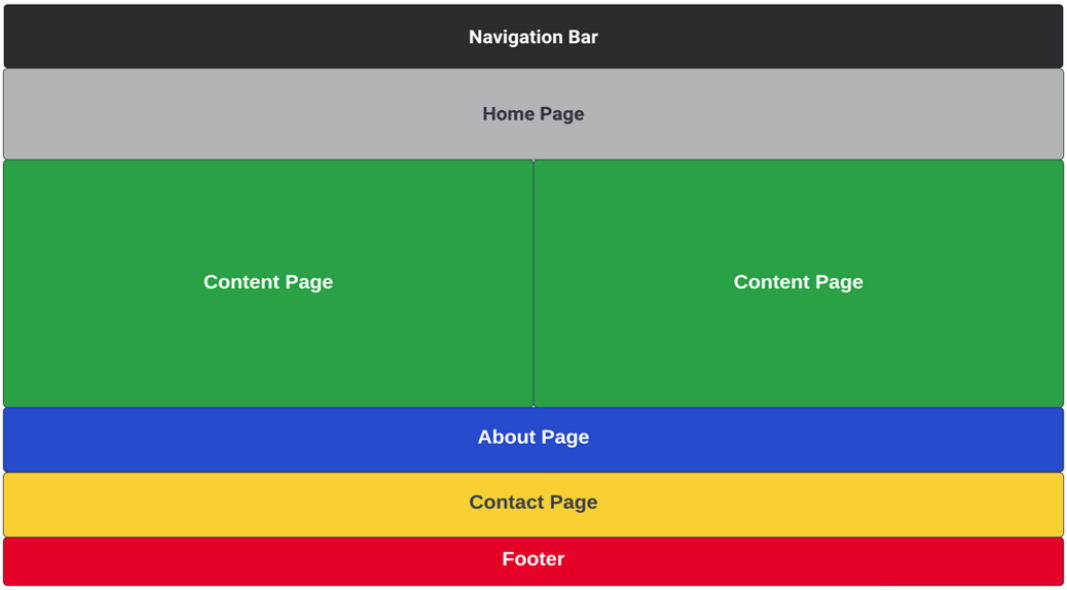


Figure 3 - User Interface Mockup of Website

5.5. Development Environment and Implementation Language

The project will leverage Python as the primary programming language due to its user-friendly syntax, extensive library support for scientific computations, data manipulation, and data visualization. Python's versatility makes it well-suited for tasks such as parsing XML data, processing datasets, and creating insightful visualizations.

The chosen development environment for this project is Visual Studio Code (VS Code). VS Code offers a robust integrated development environment (IDE) that enhances the Python development experience with features like code highlighting, debugging, and seamless integration with version control systems. It provides a flexible and efficient workspace for coding, testing, and debugging Python scripts.

5.6. Design Workflow

1. **Data Extraction:** Collecting data from the DBLP database forms the basis for all subsequent analysis.
2. **Data Parsing and Cleaning:** The extracted data is processed and cleaned, ensuring it is in a format ready for analysis.

3. **Gender Classification:** Authors' genders are inferred from their names in this step. This might require the use of external APIs or custom machine learning models.
4. **Data Analysis:** The prepared data is then analysed to identify trends in gender representation.
5. **Data Visualization:** Findings from the analysis are represented visually, aiding in interpreting, and understanding the results.
6. **Web Application Development:** The visualizations are incorporated into a user-friendly web application that allows users to explore the data.
7. **Testing and Iteration:** The entire process is continually tested and refined, with issues addressed as they arise.
8. **Final Deployment:** Once testing is completed and the system is robust, the final product is deployed.

This approach ensures a systematic, sequential workflow for the project, with each stage building on the last, promoting efficient project realization.

Chapter 6. IMPLEMENTATION

During the implementation of the project, I made several important decisions and changes to the original plan. Here is a summary of those changes:

6.1. Changes Made During Implementation

1. **Data Retrieval:** Initially, I had planned to connect directly to the DBLP database for data retrieval. However, I switched to using the DBLP API, which proved to be a more streamlined and efficient way to fetch up-to-date data.
2. **Data Parsing:** While the initial plan involved creating a complex XML parser for the raw data from the DBLP API, I discovered that the API's response was already well-structured, this made data parsing simpler, eliminating the need for a complex parser.
3. **Gender Classification:** For author names' gender classification, I initially considered using either a Gender API or a custom machine learning model. However, in the final implementation, I chose to use a Kaggle dataset for this purpose. This dataset provided a comprehensive list of author names with associated genders, simplifying the gender classification process without the need for external APIs or custom models.

4. **Data Analysis:** The identification of trends in gender representation in computer science research involved data analysis using Python's Pandas, NumPy libraries, as planned.

5. **Data Visualization:** Visualizations of the identified trends were created as intended, utilizing Matplotlib and Plotly to represent the analysed data effectively.

6. **Web Application:** As part of the implementation, a user-friendly web application was successfully developed to enable users to interact with the visualized data and findings, as outlined in the original design.

Overall, these changes in the implementation process ensured a more efficient and effective workflow, simplifying data retrieval and gender classification while maintaining the core aspects of data analysis and visualization as initially planned.

6.2. Implementation workflow

6.2.1. Data Extraction and Processing

- **DBLP Data Extraction :** I started by getting data from the DBLP API, which provides information about computer science publications. I made requests to the API based on vowel queries (a, e, i, o, u) to gather details like publication titles, authors, years, venues, and more. It was a bit challenging because the data came in a nested XML format, which required careful parsing.

6.2.2. **Kaggle Dataset Download :** I also downloaded a Kaggle dataset that predicts gender based on names. This dataset was essential for classifying authors by gender based on their first names. However, I needed to ensure that the dataset had accurate gender predictions.

6.2.3. **Data Processing and Enrichment :** After gathering the data, I processed it by converting string lists into actual Python lists for better handling. I also classified author genders using the Kaggle dataset and prepared the data for visualization and analysis.

6.2.4. Data Visualization

I created various visualizations to explore gender diversity trends:

- **Gender Distribution Analysis:** An interactive donut chart showed the overall gender distribution.

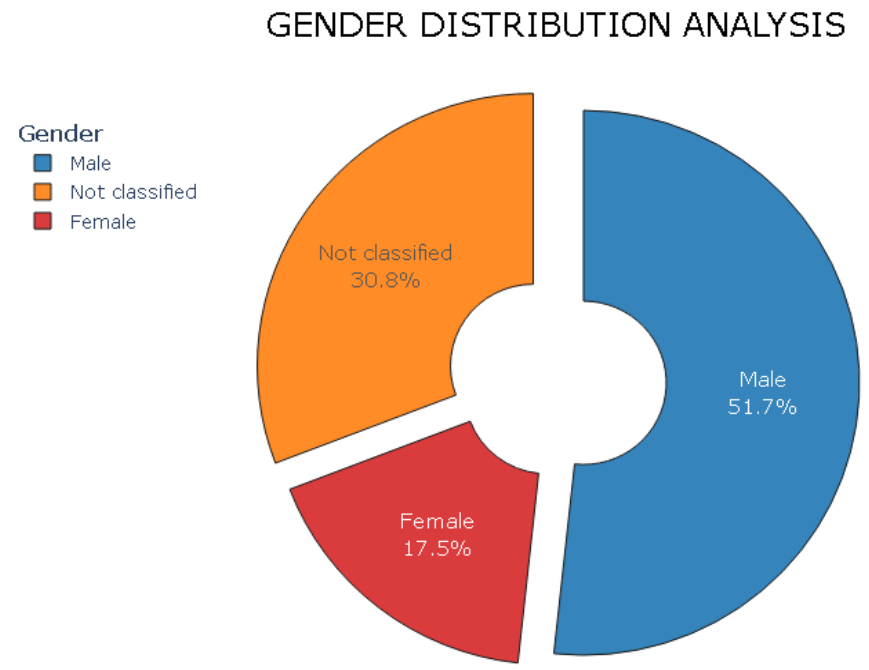


Figure 4 - Interactive Donut for Gender Distribution Analysis

- **Yearly Publication Count with Gender Trend Lines:** A line plot showing the trend of publication counts over the years for different gender categories.

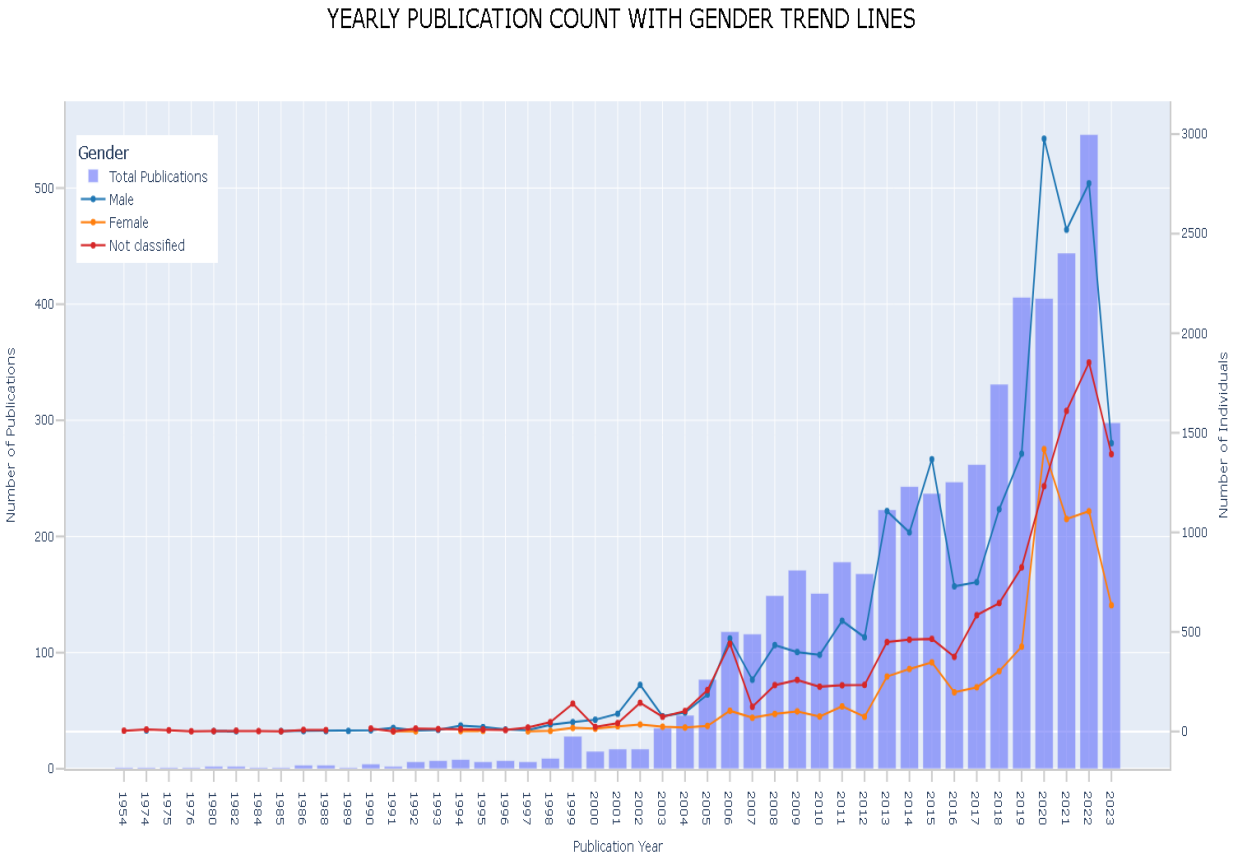


Figure 5 - Interactive Plot for Yearly Publication Count with Gender Trend Lines

- **Publication Venue Distribution Over Years:** This visualization displayed the distribution of publications across different venues.

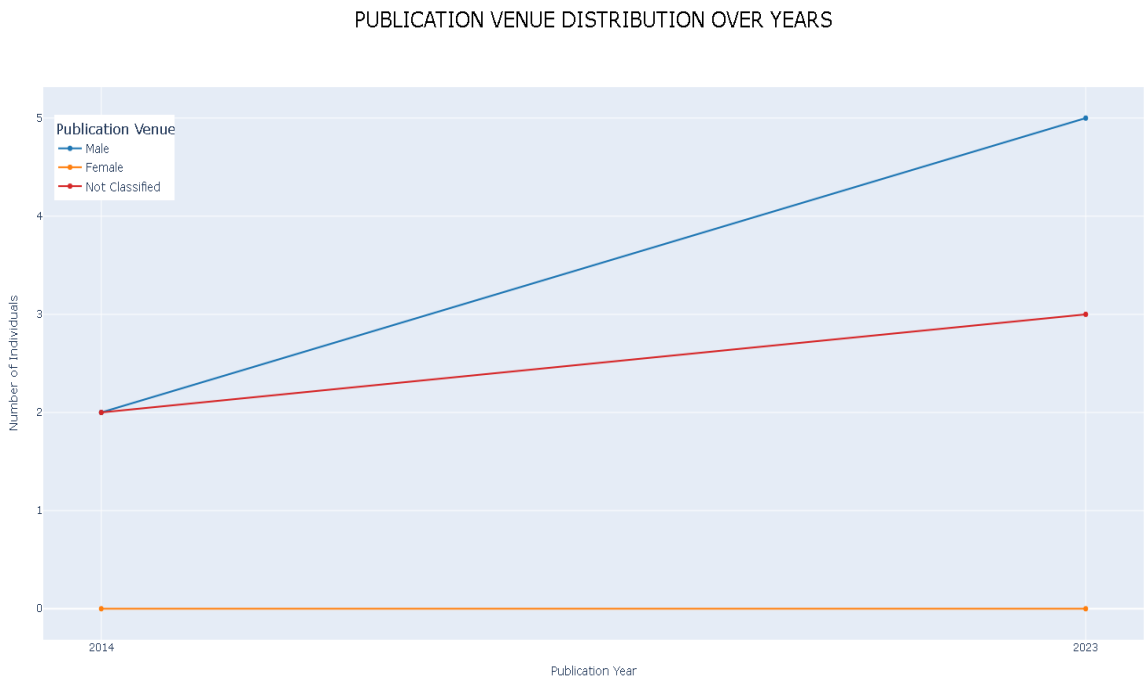


Figure 6 - Interactive Plot for Publication Venue Distribution Over Years

- **Publication Type Distribution Over Years:** Like venue distribution, this one explored publication types.

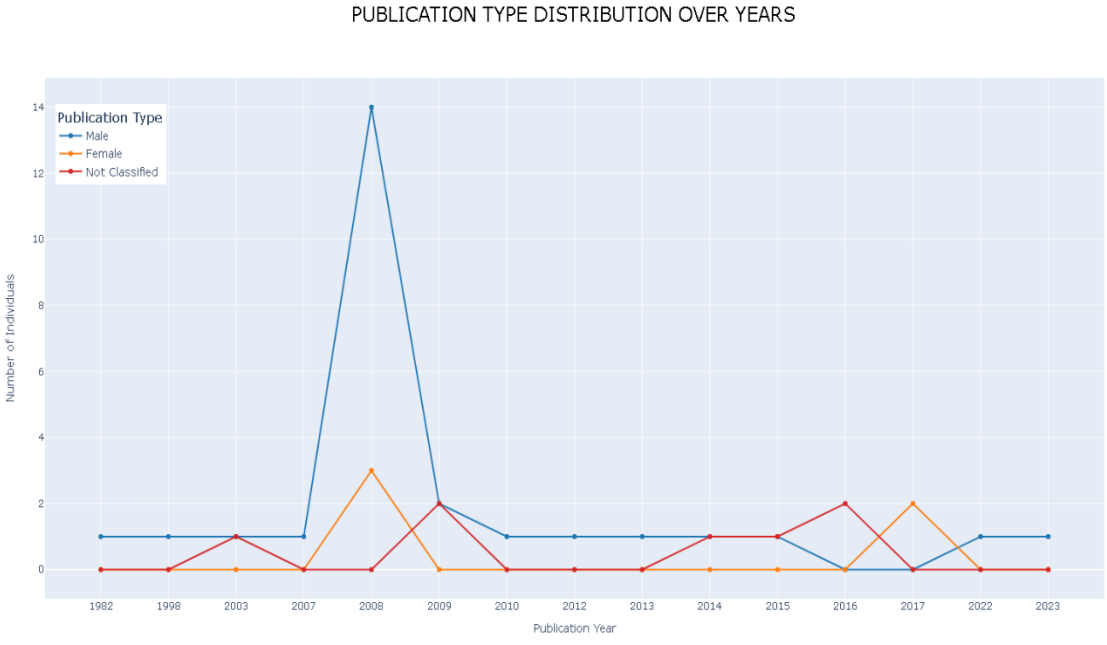


Figure 7 - Interactive Plot for Publication Type Distribution Over Years

6.2.5. Web Application Development

The core of the project was the development of an interactive web application using the Dash framework. The web application included the following components:

- Navigation bar for easy access to different sections of the application.
- Home page with a brief introduction to the project.
- Gender Distribution section displaying the donut chart.
- Yearly Gender Trend section displaying the trend lines.
- Publication Venue & Type Distribution section with a dropdown for venue & type selection.
- About page with project information, contact page for inquiries.

6.3. Testing

I rigorously tested each component:

- I checked data processing functions to ensure data transformations and classifications were accurate.
- I evaluated visualizations for correctness and interactivity.
- I thoroughly tested the web application for navigation, responsiveness, and functionality.

6.4. Achievements

I successfully extracted, processed, and enriched data from both the DBLP API and the Kaggle dataset. I designed interactive data visualizations and developed a user-friendly web application with a responsive interface. I also maintained data quality control for accurate gender predictions based on first names.

6.5. Challenges and Dead Ends

- Parsing and extracting data from nested XML responses from the DBLP API was challenging and time-consuming. Handling missing or incomplete data required careful consideration.

- Some names in the Kaggle dataset were not classified, and the accuracy of gender predictions was not 100%. Dealing with unclassified names and ensuring accurate predictions was an ongoing challenge.

The implementation of project involved various technical challenges, data processing, visualization, and web application development. The project aimed to provide a comprehensive view of gender diversity in computer science publications over time and make the findings accessible through an intuitive web interface. Achievements and challenges in each phase of implementation were critical in delivering the final product.

Chapter 7. EVALUATION

Strengths:

1. **Effective Data Visualization:** The project excels in creating charts and graphs that make it easy to understand gender diversity trends in computer science publications.

2. **User-Friendly Web App:** The user interface of the web application is intuitive, allowing users to explore the data effortlessly.

3. **Ethical Data Use:** The project uses publicly available and anonymized data sources, ensuring privacy and transparency.

4. **Adaptability:** The project successfully adjusted to different data sources, improving data retrieval efficiency.

Weaknesses:

1. **Gender Classification Accuracy:** The project's gender predictions based on names may not always be accurate.

2. **Limited Scope:** It primarily focuses on gender diversity and may not consider other crucial factors like race or socioeconomic background.

3. **Data Source Challenges:** Dealing with complex data sources and missing data was a challenge.

4. **Scope:** The project's findings may not apply to other STEM fields or disciplines.

Feedback-Driven Changes:

I switched from XML parsing to using the DBLP API to streamline data retrieval based on feedback and challenges faced. Ethical data practices remained a core principle.

In summary, the project excels in data visualization and user-friendliness but faces challenges related to gender prediction accuracy and limited scope. Adjustments were made to improve efficiency, and the project contributes to discussions on gender diversity in computer science publications.

Chapter 8. LEARNING POINTS

8.1. Key Learning Points

1. **Efficient Data Retrieval:** Getting data efficiently is crucial. Using the DBLP API was more efficient than dealing with complex data formats.

2. **Data Ethics:** I learned to use data responsibly. Public and anonymized data sources protect privacy and ensure transparency.

3. **Adaptability:** Being open to change can make projects smoother. I improved the methods during the project.

4. **Gender Prediction Challenges:** Predicting gender based on names can be tricky and not always accurate. I learned to understand these limitations.

5. **User-Friendly Design:** Making the web app user-friendly is important for accessibility and impact.

6. **Managing Complexity:** Handling complex data and addressing issues, like missing data, was a valuable skill.

7. **Interdisciplinary Learning:** I saw how different fields like data science, ethics, and sociology are connected in addressing gender diversity in computer science.

8.2. Crucial Actions for Success

- **Listening to Feedback:** Taking feedback from users and collaborators helps improve the project.

- **Thorough Testing:** Testing at every step makes sure the project works correctly.

- **Transparency:** Being clear about data sources and ethics is essential.

8.3. Future Improvements

- **Better Gender Prediction:** Exploring more accurate ways to predict gender beyond just names.
- **Inclusivity:** Expanding the project to consider factors like race, ethnicity, and economic background.
- **Collaboration:** Working with experts from various fields can bring new perspectives.

In short, I learned about efficient data handling, ethical data use, adaptability, user-friendly design, complexity management, and interdisciplinary thinking. Feedback, testing, and transparency are crucial for success. Future improvements may involve better gender prediction, inclusivity, and collaboration.

Chapter 9. PROFESSIONAL ISSUES

My project aligns with the British Computer Society (BCS) Code of Conduct, adhering to ethical and professional standards. Here is how:

1. **Public Benefit:** My project aims to benefit the public by addressing gender diversity issues in computer science and making the findings accessible through a user-friendly web app.
2. **Competence and Integrity:** I maintained professionalism by using ethical data sources, ensuring data privacy, and carefully selecting gender prediction methods for fairness.
3. **Responsibility:** I followed ethical guidelines and respected data source policies, ensuring authorized access to data.
4. **Public Duty:** My project serves the public interest by promoting diversity and inclusion in computer science.
5. **Professional Contribution:** I contribute to the profession by advancing knowledge and sharing research findings.

Regarding data, I used publicly available sources and anonymized gender prediction based on names, complying with ethical and privacy standards. My project did not involve human participants, surveys, or sensitive personal data.

In summary, my project aligns with the BCS Code of Conduct, follows ethical data practices, and does not involve human participants or sensitive data.

Chapter 10. CONCLUSIONS

In this project, I set out to explore the history of gender diversity in computer science publications and develop a user-friendly website to visualize the findings. I aimed to answer questions about the representation of genders in this field and understand how it has evolved over time.

10.1. What I Achieved

1. **Data Collection**: I gathered data on computer science publications and inferred the gender of authors based on their names. This involved using the DBLP API for publication data and a Kaggle dataset for gender predictions.

2. **Data Visualization**: I created various charts and graphs to present the data effectively, allowing for insights into gender diversity trends.

3. **Web Application**: I developed an interactive website that enables users to explore the research findings through different sections, including gender distribution and trends.

10.2. Key Findings

- While men still dominate computer science publications, there has been a gradual increase in women's participation over the years.

- Gender diversity varies across different publication venues, such as conferences and journals.

- The type of publication, whether conference papers or journal articles, has an impact on gender diversity, with conference papers showing higher inclusivity.

10.3. Future Work

1. **Intersectionality**: Future research can delve into how gender intersects with other factors like race and background to gain a more comprehensive understanding.

2. **Ethical Considerations**: Continuing discussions on the ethics of inferring gender from names in research is crucial.

3. **Global Perspective**: Examining gender diversity on a global scale can shed light on regional variations.

4. **Interventions**: Implementing policies and initiatives to enhance gender diversity in computer science remains a priority.

In conclusion, this project offers insights into the historical gender diversity landscape in computer science. While progress has been made, there is still work to be done to create a more inclusive and equitable environment in this field.

BIBLIOGRAPHY

1. Ley, M. (2009). DBLP. *Proceedings of the VLDB Endowment*, 2(2), pp.1493–1500. doi: https://doi.org/10.14778/1687553.1687577.

2. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M. and Strohmaier, M. (2016). *Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods*. [online] arXiv.org. doi: https://doi.org/10.1145/2872518.2889385.

3. Vasilescu, B., Posnett, D., Ray, B., van den Brand, M.G.J., Serebrenik, A., Devanbu, P. and Filkov, V. (2015). Gender and Tenure Diversity in GitHub Teams. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. doi: https://doi.org/10.1145/2702123.2702549.

4. Correll, Shelley J. (2001). Gender and the Career Choice Process: The Role of Biased Self-Assessments. *American Journal of Sociology*, [online] 106(6), pp.1691–1730. doi: https://doi.org/10.1086/321299.

5. Wang, M.-T. and Degol, J.L. (2016). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, 29(1), pp.119–140. doi: https://doi.org/10.1007/s10648-015-9355-x.

6. West, J.D., Jacquet, J., King, M.M., Correll, S.J. and Bergstrom, C.T. (2013). The Role of Gender in Scholarly Authorship. *PLoS ONE*, 8(7), p.e66212. doi: https://doi.org/10.1371/journal.pone.0066212.

7. Holman, L., Stuart-Fox, D., and Hauser, C.E. (2018). The gender gap in science: How long until women are equally represented? *PLOS Biology*, [online] 16(4), p.e2004956. doi: https://doi.org/10.1371/journal.pbio.2004956.

8. Age-of-the-sage.org. (2019). *Social Identity Theory - Tajfel and Turner 1979*. [online] Available at: https://www.age-of-the-sage.org/psychology/social/social_identity_theory.html.

9. Anon, (n.d.). *UAL Creative Mindsets*. [online] Available at: https://ualcreativemindsets.myblog.arts.ac.uk/stereotype-threat/.

10. William & Mary. (n.d.). *Inclusive Excellence*. [online] Available at: https://www.wm.edu/offices/diversity/inclusive-excellence/.

11. Shah, H.S. and Bohlen, J. (2023). *Implicit Bias*. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK589697/.

12. Dickey, M.R. (2021). *Examining the 'pipeline problem'*. [online] TechCrunch. Available at: https://techcrunch.com/2021/02/14/examining-the-pipeline-problem/.

13. Mindset Works (2017). *The Growth Mindset - What is Growth Mindset - Mindset Works*. [online] Mindsetworks.com. Available at: https://www.mindsetworks.com/science/.

14. Dblp.org. (2023). Available at: https://dblp.org/search/publ/api

15. www.kaggle.com. (n.d.). *Gender Prediction by using names*. [online] Available at: https://www.kaggle.com/datasets/monukhan/gender-prediction-by-using-name.

APPENDICES

Appendix A: Project Log

Activity	Timeline
Project Initiation	June 5, 2023
Background Reading and Literature Review	June 12, 2023 - June 16, 2023
Development of Project Specification and Proposed Design	June 19, 2023 - July 14, 2023
Development of Project Specification and Proposed Design Initial Presentation Submission Deadline	July 14, 2023
Data Collection and Processing	June 15, 2023 - July 10, 2023
Software Implementation and Testing	July 17, 2023 - August 25, 2023
Meeting with Supervisor regarding Gender Identification Challenge	August 11, 2023
Web Application Development	August 14, 2023 - August 18, 2023
Data Visualization	August 21, 2023 - August 25, 2023
Testing and Debugging	August 28, 2023 - September 1, 2023
Software Implementation and Testing Final Presentation Submission Deadline	September 1, 2023
Write-Up of Dissertation	September 4, 2023 - September 22, 2023
Q&A Meeting	September 7, 2023
Dissertation Final Submission Deadline	September 22, 2023

Appendix B: Python Version and Libraries Used

Python Version: The project was developed using Python 3.11.3, leveraging the latest features and enhancements available in this version.

Libraries Used:

- Requests
- xml.etree.ElementTree
- Pandas
- Plotly and Plotly Express
- Dash
- dash_bootstrap_components
- Kaggle API
- zipfile
- NetworkX

These libraries were instrumental in various aspects of the project, from data retrieval and manipulation to interactive visualization and web application development.

Appendix C: Web Application

The hyperlink to the web application is generated after the successful execution of the code. Below is the screenshot, of the web application.

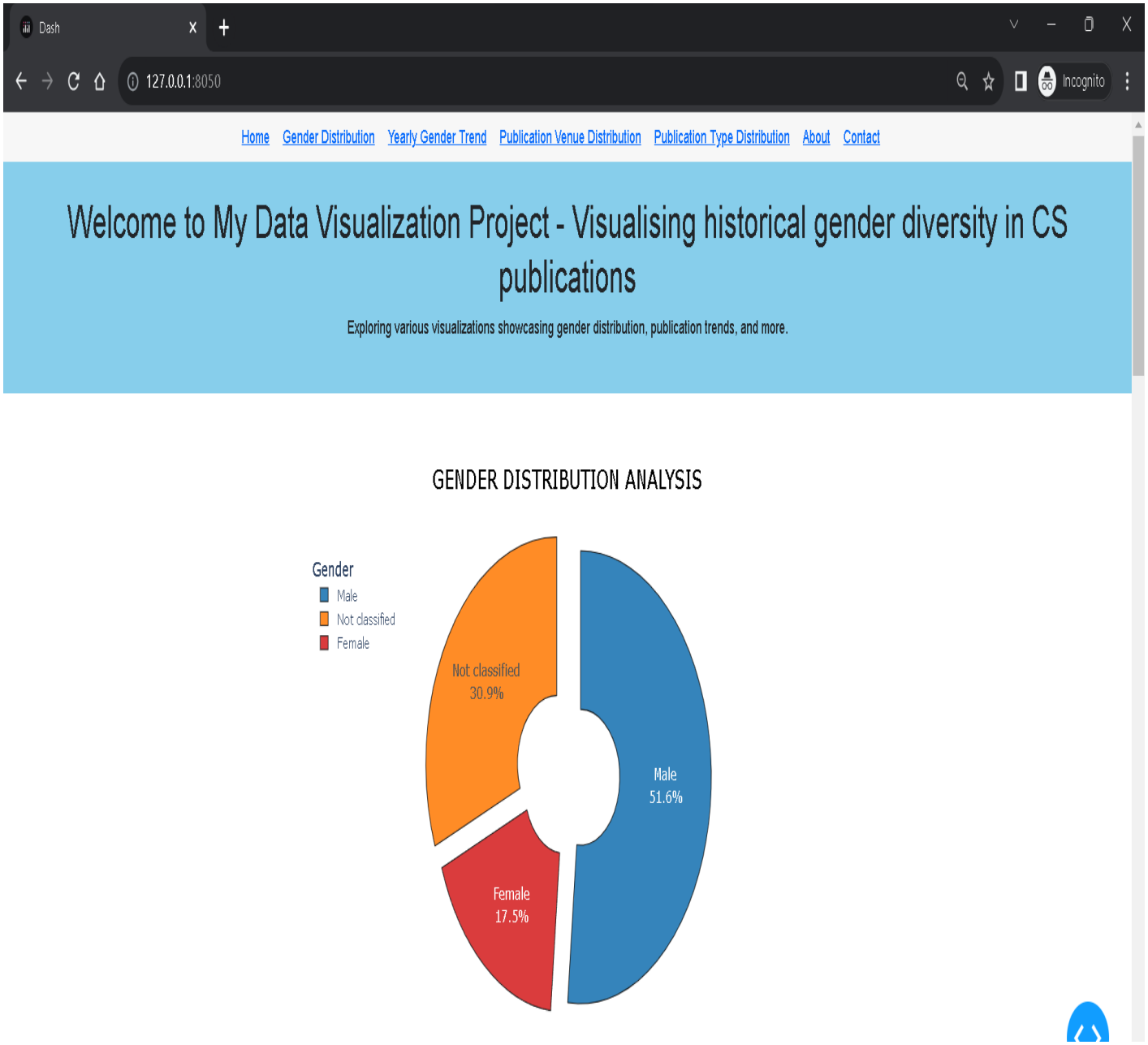


Figure 8 - Sample Image of Website Created