SANTAR: Swimming Assessment using Numerical daTa Analytics and visual Representation

Nishtha Agarwaal
School of Computing and Data Sciences, FLAME University

Defense Date: 20 April 2024

Thesis Advisor:
Chiranjoy Chattopadhyay, Ph.D.
School of Computing and Data Sciences, FLAME University

Defense Committee:
Chiranjoy Chattopadhyay, Ph.D.
School of Computing and Data Sciences, FLAME University
Professor Jayaraman V. K.
School of Computing and Data Sciences, FLAME University

1 Abstract

This report explores SANTAR (Swimming Assessment using Numerical Data Analytics and Visual Representation), a project leveraging data analysis to understand competitive swimming performance. Data from World Championships (2000-2023) was collected and cleaned, focusing on metrics like rank, time, and points. Tableau software transformed the data into eight key visualizations, including world maps depicting country dominance, heatmaps revealing age-related trends, and circle charts analyzing lane wins/losses. These visualizations provide valuable insights for coaches, athletes, and swimming enthusiasts. Coaches can utilize them to tailor training plans and benchmark performance. Athletes gain a deeper understanding of their strengths and weaknesses. The report acknowledges challenges like data collection time and ethical considerations. Future directions include expanding the data scope, incorporating biomechanical data, and exploring machine learning for performance prediction. By harnessing data analysis, SANTAR empowers stakeholders to optimize training and propel swimmers toward achieving their full potential.

2 Introduction

The world of competitive swimming is a dynamic one, where fractions of seconds can separate victory from defeat. Swimmers, coaches, and fans alike are constantly seeking new ways to gain an edge, to understand the intricate factors that contribute to peak performance. Traditional methods of analyzing swimming data often rely on visual inspection, providing valuable insights but potentially overlooking hidden patterns and trends buried within the vast amount of information available.

SANTAR (Swimming Assessment using Numerical Data Analytics and Visual Representation) is a data mining and visualization tool designed to delve deeper and explore and analyze the trends within historical swimming data. This project, named after the Bengali word for "swimming," transcends the limitations of traditional analysis by using numerical data exploration. SANTAR leverages data from the FINA Swimming World Championships held between the years 2000 and 2023, offering a comprehensive dataset ripe for uncovering the nuanced factors that influence competitive swimming.

The significance of SANTAR lies in its ability to provide a more holistic understanding of the sport. By employing sophisticated data mining techniques, SANTAR allows the user to analyze one's performance based on several factors and also makes some interesting observations and predictions. This deeper analysis holds tremendous potential to benefit a wide range of stakeholders within the swimming community. Aspiring swimmers can gain valuable insights into their plans to get on track to the World Championships. For professional athletes, SANTAR can offer personalized visualizations, enabling them to tailor their training to specific stroke weaknesses or optimize race tactics based on historical trends. Coaches gain a powerful tool to analyze individual and team performance, creating data-driven training plans that maximize potential. Even swimming enthusiasts and fans

can delve into the rich tapestry of data, appreciating the intricate factors that contribute to world-class performance.

My contribution to this innovative project centers around the development of user-friendly dashboards using Tableau, a powerful data visualization software. These dashboards translate complex swimming data into clear and concise visuals, allowing users to explore trends, identify patterns, and gain a deeper understanding of the information at hand. Furthermore, I have formulated a central research question that serves as the driving force behind SANTAR: "Can past results data be analyzed to identify patterns and trends that can inform swimmer training, team strategies, and our overall understanding of competitive swimming?"

SANTAR's investigative power extends to eight key areas, each designed to shed light on a specific aspect of swimming performance. Firstly, the project delves into country performance analysis, revealing historical trends and identifying nations that consistently dominate specific strokes or distances. Secondly, SANTAR focuses on individual swimmer performance, providing a comprehensive assessment of an athlete's performance across different competitions over the years. The project goes further by comparing individual performance between heats and finals, uncovering factors that influence an athlete's ability to peak at the right moment. Understanding how an individual performs across different strokes, such as butterfly, freestyle, or backstroke, is crucial, and SANTAR offers an analysis in this area.

Furthermore, the project investigates an individual's performance across various distances within a stroke. For example, analyzing performance in the 50-meter freestyle compared to the 200-meter freestyle can reveal an athlete's strengths in speed versus endurance. SANTAR empowers users to conduct detailed comparisons between swimmers based on three user-defined factors, enabling coaches and athletes to identify key areas of differentiation between competitors. SANTAR also explores the fascinating relationship between age and swimming performance, uncovering potential peak years for swimmers. Finally, the project delves into a captivating question: what is the probability of outside lane swimmers winning finals based on historical data? By analyzing past results, SANTAR can provide valuable insights into potential race strategies and psychological factors that might influence performance when starting from a non-central lane.

SANTAR stands as a testament to the power of data analysis in revolutionizing our understanding of competitive swimming. Through its innovative approach, the project aims to unlock valuable insights that can transform training methods, optimize team strategies, and elevate our overall comprehension of the sport. As we dive deeper into the data, we unlock the secrets that lie beneath the surface, propelling competitive swimming towards a new era of data-driven performance and excellence.

3 Literature Review

For decades, the quest to understand and optimize swimming performance has captivated coaches, athletes, and researchers alike. Traditionally, visual inspection of race footage and competition results formed the cornerstone of performance analysis [1]. While this approach offers valuable insights into stroke mechanics, race tactics, and overall competitiveness, it can overlook hidden patterns and trends buried within the vast amount of data available. Recognizing this limitation, recent years have witnessed a surge in data-driven analysis methods aimed at extracting deeper and more nuanced understandings of swimming performance.

Pioneering work by [2] explored the application of statistical analysis to race data from the International Swimming Federation (FINA) World Championships. Their focus centered on quantifying race strategies and characteristics across different strokes and distances (100m, 200m, etc.). This groundbreaking approach demonstrated the potential of data analysis to provide objective information beyond what the human eye can readily perceive. For instance, [2] were able to identify subtle differences in pacing strategies between male and female swimmers in the 1500m freestyle, highlighting the potential for data analysis to reveal gender-specific performance nuances. Building upon this foundation, other researchers have delved deeper into specific aspects of performance. For example,[3] employed regression analysis, a statistical technique, to identify the relationship between age, gender, and anthropometric factors (body composition measurements) on swimming performance in elite athletes. Their findings shed light on the complex interplay between physical attributes and competitive success, revealing that factors like height, wingspan, and body fat percentage can influence performance across different strokes [3].

Despite these advancements, existing solutions often exhibit limitations. Some solutions focus on relatively narrow aspects of performance, such as race strategy [2] or specific biomechanical factors like underwater dolphin kick efficiency [4]. While valuable, these targeted approaches lack the comprehensive nature required for a holistic understanding of performance. Additionally, some existing solutions may utilize complex statistical techniques that require advanced expertise, limiting their accessibility to coaches and athletes who lack a strong statistical background [1].

SANTAR builds upon the foundation laid by previous research but with several key differentiators. Firstly, it utilizes a wider range of data points, encompassing World Championship results from 2000 to 2023. This comprehensive dataset, spanning over two decades, allows for a more robust analysis, capturing a broader spectrum of performance trends and athlete characteristics. For instance, SANTAR can be used to investigate historical dominance patterns in specific strokes by particular countries, potentially revealing training philosophies or cultural influences that contribute to success. Secondly, SANTAR leverages the power of data visualization through user-friendly dashboards. These dashboards translate complex data into clear visuals, such as interactive charts and graphs, making the information readily accessible to a wider audience, including coaches, athletes, and even

fans. This user-centric approach empowers users to explore trends, identify patterns, and gain deeper insights without requiring advanced statistical knowledge.

Furthermore, SANTAR investigates a broader range of factors influencing performance, venturing beyond the limitations of previous studies. It delves into areas like country analysis, uncovering historical trends and identifying nations that consistently dominate specific strokes or distances. Additionally, SANTAR focuses on individual performance across strokes and distances, allowing coaches and athletes to tailor training plans based on an athlete's strengths and weaknesses across different disciplines. The project also explores the intriguing question of lane position's impact on success. By analyzing historical data on lane assignments and final placements, SANTAR can provide valuable insights into potential race strategies and psychological factors that might influence performance when starting from a non-central lane. For instance, SANTAR might reveal trends suggesting that athletes in outer lanes adopt more aggressive race tactics to compensate for the perceived disadvantage.

SANTAR doesn't simply propose another solution; it aspires to revolutionize the way we analyze and interpret swimming data. By offering a user-friendly, comprehensive, and data-driven approach, SANTAR empowers a wider range of stakeholders within the swimming community to gain deeper insights and optimize performance. Coaches can leverage SANTAR's data visualizations and analysis to identify areas for improvement in their athletes' training regimens. Athletes can utilize the platform to understand their strengths and weaknesses, track progress over time, and potentially identify areas where data-driven training adjustments can yield performance gains. Even fans can gain a deeper appreciation for the complexities of competitive swimming by exploring the fascinating trends and patterns revealed by SANTAR's analysis. Imagine a coach being able to compare their athlete's individual performance data across different strokes and World Championships throughout their career. SANTAR could reveal hidden trends, such as a gradual decline in freestyle speed but a sustained improvement in backstroke performance. This information could prompt the coach to adjust training strategies to address the decline in freestyle speed while capitalizing on the strengths developing in backstroke.

In conclusion, the reviewed literature highlights the growing influence of data analysis in sports, particularly swimming. While existing solutions offer valuable insights, they often lack comprehensiveness, user-friendliness, or the scope to investigate a broader range of performance-influencing factors. SANTAR, with its user-centric data visualization tools and comprehensive data analysis capabilities, positions itself not just as another solution, but as a potential game-changer in the world of swimming performance analysis. By harnessing the power of data and making it accessible to a wider audience, SANTAR aims to unlock valuable knowledge and resources, ultimately propelling competitive swimming towards a new era of data-driven performance optimization. As SANTAR continues to evolve and integrate new data sources, it has the potential to revolutionize the way athletes, coaches, and fans understand and appreciate the intricacies of peak performance in the pool.

4 SANTAR: Unveiling Depths of Performance in Swimming

For decades, the quest to understand and optimize swimming performance has captivated coaches, athletes, and researchers alike. Traditionally, analyzing race footage and results formed the cornerstone of performance evaluation. While valuable, this approach can overlook hidden patterns within the vast amount of data available.

SANTAR emerges as a data-driven solution to address this limitation. It offers a comprehensive approach to understanding and optimizing swimming performance through:

1. **Data Comprehensiveness:** SANTAR utilizes a wider range of data points, encompassing World Championship results from 2000 to 2023. This extensive dataset allows for robust analysis, capturing broader performance trends and athlete characteristics.

2. **User-Friendly Data Visualization:** SANTAR translates complex data into clear visuals through interactive dashboards. This empowers users of all backgrounds to explore trends, identify patterns, and gain deeper insights without needing advanced statistical knowledge.

3. **Broader Scope of Analysis:** SANTAR delves deeper than existing solutions by investigating:

    - Country Analysis: Uncovering historical trends and identifying nations that consistently dominate specific strokes or distances.
    - Individual Performance: Analyzing performance across strokes and distances, and also their performance in heats versus finals, to tailor training plans based on an athlete's strengths and weaknesses.
    - Age and Performance: Analyzing and observing whether there is an age at which swimmers peak, taking into account the various factors that might affect this.
    - Swimmer comparison: Exploring how an individual performed against another swimmer to help tailor the swimmer's strategies and training regime.
    - Lane Position Impact: Exploring the influence of lane position in the finals on race strategy and potential psychological factors affecting performance.

By leveraging these functionalities and using 8 interactive dashboards, combined in a single Tableau story, SANTAR empowers a wider range of stakeholders within the swimming community to gain valuable insights and optimize performance.

5 Project Implementation

This section details the project's journey, outlining the steps taken, challenges encountered, limitations, and future directions.

5.1 Data Collection Strategy: Targeting World Championships for Insights

Competition Focus - World Championships were chosen as the primary data source due to several factors:

- **Frequency:** Occurring on an average of six times every year, World Championships provide a consistent data point for analysis across a significant timeframe.

- **Significance:** World Championships serve as a crucial stepping stone for Olympic preparation. Swimmers utilize these events to test strategies and refine techniques before the pinnacle event.

- **Timeframe:** The data collection process focused on the period between 2000 and 2023, capturing a comprehensive historical snapshot of World Championship performance.

5.2 Building a Structured Data Repository

Year-based Organization - To maintain clarity and facilitate future analysis, data for each year is stored in a dedicated folder locally. This ensures easy retrieval and organization based on the competition year.

- **Event-Specific Workbooks:**Within each year's folder, a separate Excel workbook is created for each World Championship stop. This separation allows for a focused analysis of individual competitions.

- **Event Sheets:** Each workbook further organizes data by event. Dedicated worksheets are created for each swimming event (e.g., Men's 100m Freestyle, Women's 200m Backstroke). This granular organization allows for in-depth exploration of performance trends within specific events. The worksheets also alternate between the finals of an event and its heat summary.

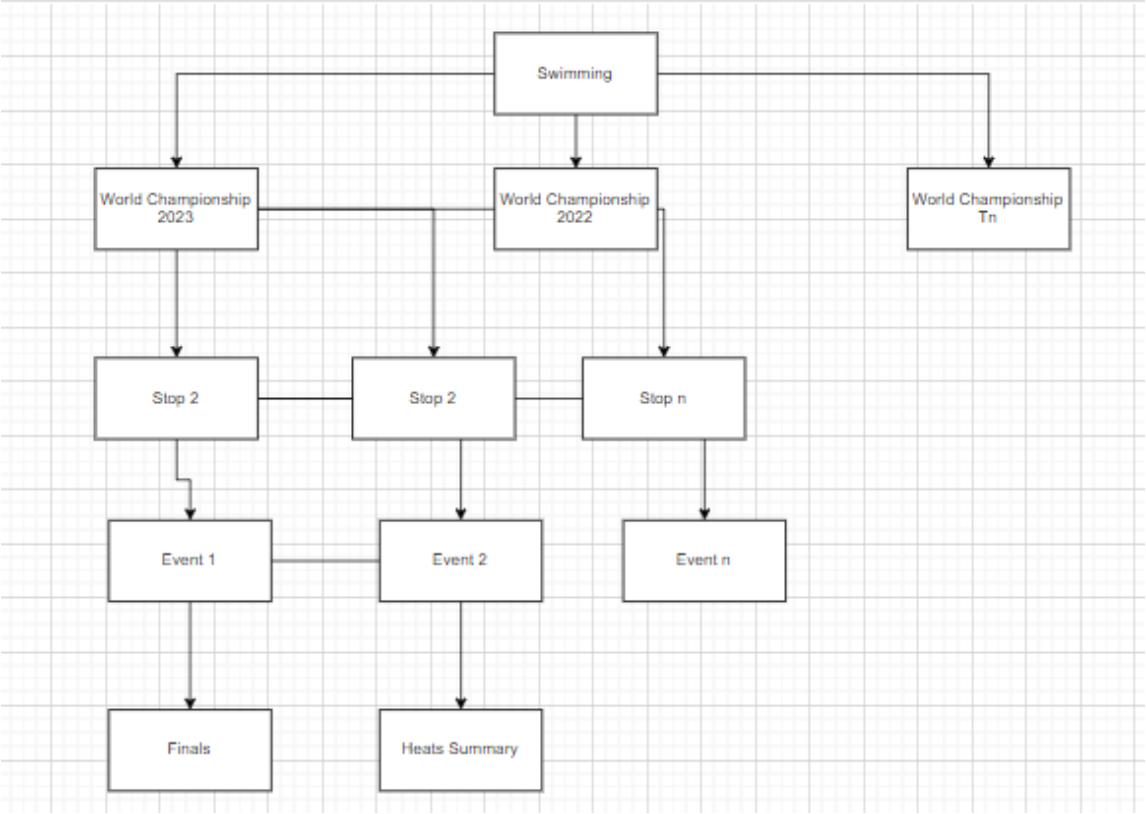

Figure 1: After analyzing the website from which data was to be scraped, A basic outline of how data was to be stored was created. The image shows the file structure and the folder structure and hierarchy in which data was stored.

5.3 Web Scraping with Python: Extracting Data from the World Aquatics Website [5]

The local repository, mentioned in the structure above, is already created manually by adding folders. Python is employed for web scraping. The code is written in a way that extracts one competition's data and stores it in a workbook. The workbook name is a manual input in the structure 'stop(no.)-(country)-year', for example 'stop1-ger-2023'. The scrapped workbooks were then manually stored in their respective year folders.

Libraries: Python, a versatile programming language, is employed for web scraping. Libraries like `selenium` for web driver using Microsoft Edge that helps in retrieving data, while `BeautifulSoup` (or `lxml`) is used to parse the retrieved HTML content, allowing us to navigate the website structure and extract specific data points. A structure for the table was created by putting the column names as given in the website and data was stored

accordingly.

Target Website: Data is extracted from the official World Aquatics website (the specific URL of a competition stop would be required to implement the code).

Ethical Considerations: Respectful scraping practices are paramount.

5.4 Data Cleaning: Preparing the Raw Data for Analysis

Once the data was scrapped, it underwent a cleaning process to ensure its accuracy and usability. This involved identifying and handling inconsistent formatting. Data entries exhibited some inconsistencies in formatting (e.g., variations in how time is represented, a middle name column with first name and last name for some, etc.). These inconsistencies were addressed to ensure uniformity within the dataset.

For example, there was an inconsistency in the names columns. There were two designated columns, first name and last name. While this should have been true for all rows, some swimmers had a middle name as well. This was handled manually by checking the number of columns for each row. If it was greater than 10 (the normal length), we found the column that was extra by checking the data type. The next column after the name should be time with a data type of float. The code simply ignored the third name column and added the time, in its column.

5.5 Exploring Data Visualization with Tableau

Tableau is a powerful data visualization software that was used to transform the cleaned data into clear and insightful visuals. These visuals will communicate patterns and trends within the data in an easily understandable way, empowering users to gain valuable insights without needing extensive statistical expertise. A significant amount of time was spent on learning and understanding tableau to create detailed, user-friendly visualizations.

Tableau's official website has a tutorial that walks the user through all the basic necessities to know about Tableau to get started and create visualizations [6]. The main attention was given to learning the difference between joins and relationships, creating calculated fields, changing the types of the columns specific to visualizations that needed to be created, adding filters to specific fields, and understanding what measure fields and calculated fields in Tableau (based on which Tableau shows graphs and charts available).

5.6 Finalizing Visualization Choices: Selecting the Most Effective Views

A variety of visualization techniques can be employed to represent the data. Careful consideration was given to the following while selecting the visualizations:

- **Data Type:** The chosen visualization should be appropriate for the type of data being presented (e.g., bar charts for categorical data, line charts for trends over time).

- **Clarity and Conciseness:** Visualizations should be clear, concise, and easy to interpret. Excessive complexity can hinder user understanding.

- **Audience:** The target audience's familiarity with data analysis was considered when selecting visualizations. Simpler visualizations may be preferred for a broader audience.

Attention was also paid to what filters the users might need for a more clear and concise interpretation. Relevant to the visualizations, the filter of swimmer name, month of the year, and country were provided.

5.7 Restructuring Data for Visualization Needs

While the initial data structure facilitates organization, further restructuring was necessary to optimize visualization creation in Tableau. This involved:

- **Combining Worksheets:** Data from multiple worksheets within a workbook (e.g., results from different heats of the same event) were combined into a single worksheet for a more comprehensive view.

- **Creating Calculated Fields:** New calculated fields were created within Tableau to derive additional insights from the existing data. For instance, a calculated field could be created to represent dates of different World Championships.

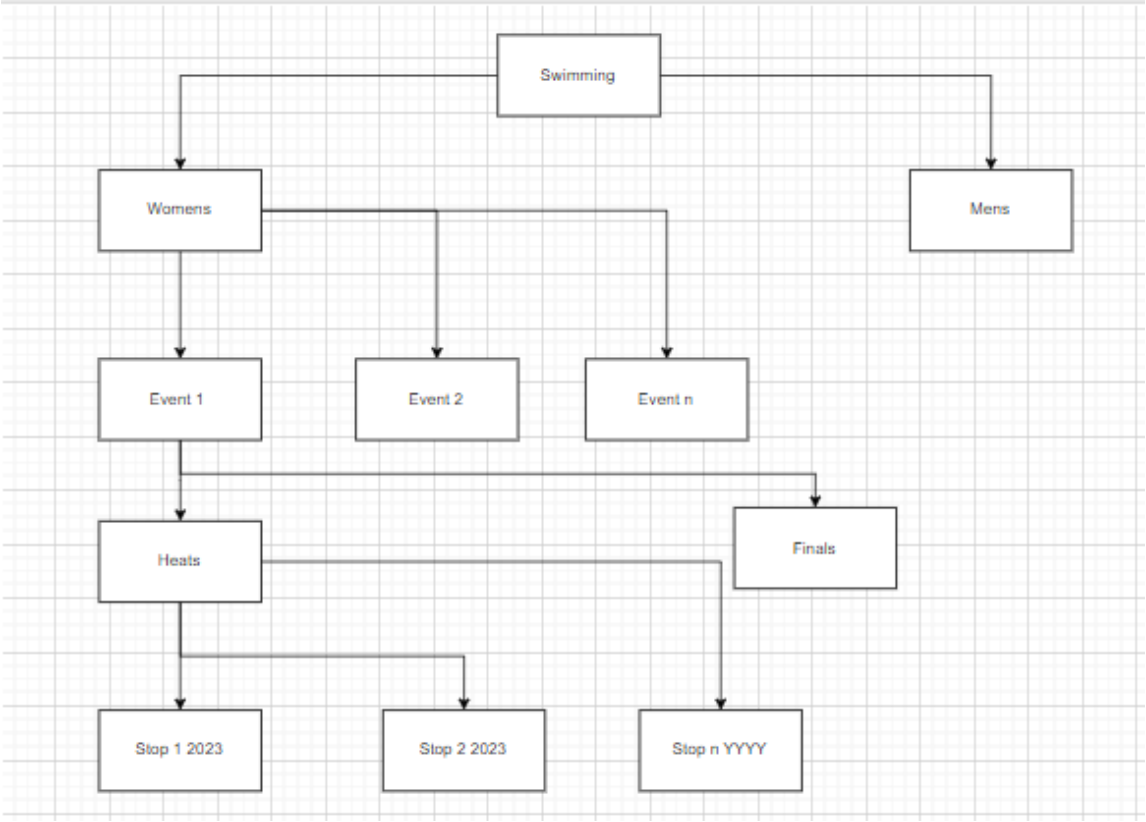

Figure 2: After evaluating the performance of joins in tableau, data was restructured in Excel using the above hierarchy and file structure for better and faster performance.

These adjustments will ensure the data is presented in a format that aligns seamlessly with the chosen visualizations.

5.8 Calculated Fields in Tableau: Deriving Additional Insights

Tableau's ability to create calculated fields empowers us to extend the analysis beyond the raw data. Here are some examples:

- **Outside Lane Wins:** A calculated field was created to count the number of times a race was one from an outside lane in the finals for a specific event. The calculation is performed on the lane column from the dataset. Function used:

  IF [Lane] IN (1,7,8,9) AND [Rank Backstroke] = 1 THEN 1

  ELSEIF [Lane] IN (1,7,8,9) AND [Rank Backstroke] = 2 THEN 1

  ELSEIF [Lane] IN (1,7,8,9) AND [Rank Backstroke] = 3 THEN 1

ELSE 0

END

This calculation creates a new column and adds it to our dataset. It checks if the value of the lane column is an outside lane and if the rank for that lane is either 1, 2, or 3, if true, the calculation writes 1 for that cell otherwise writes 0.

- **Date:** A calculated field was created to produce a date from the stop name/ file name to plot the performance over years and different months. The calculation is performed on the first column of each worksheet where the stop number, country of completion, and year are stored as follows - stop1-ger-2023. Function used:

  DATE(DATEPARSE ("dd-MM-yyyy", "01-"+ MID([Sheet], 5, 1)+"-"+RIGHT([Sheet], 4)))

  This calculation adds a new column to our dataset and converts the string field into a date-type field. The date is assumed to be 1 by default and the stop number is converted to month and the year is extracted from the string.

- **Time:** A calculated field was created to accurately rewrite the time of longer events only in seconds since a float data type can not handle more than one decimal point. This calculation is performed on the time field of the dataset where time might be in the format - 01:25:56 (min:sec:frac). To convert this into usable data, the following function was used:

  IF LEN([Time 2]) = 7 THEN

  FLOAT(MID([Time 2], 1, 1)) * 60 +

  FLOAT(MID([Time 2], 3, 2)) +

  FLOAT(MID([Time 2], 6, 2)) / 100

  ELSE

  INT(MID([Time 2],1, 2)) * 60 + INT(MID([Time 2], 4, 2))

  END

  To perform this calculation, the original time column was separated into three new columns minutes, seconds, and fractions, using the position and length that was needed. Next, the above calculation was performed which converts all the columns into seconds based on their property and adds the resulting time to create a new column called 'Time in (event)'.

These calculated fields will unlock new layers of insights within the data, enriching the overall analysis.

5.9 Visualization Creation and Refinement: Finding the Best Fit

With the data prepared and calculated fields established, the next step taken was visualization creation. Tableau offers a wide range of chart types, and the selection process involved:

- **Experimentation:** Various visualizations were created to explore which ones best represent the data and the intended message.
- **Aesthetics and Design:** Attention was paid to visual aesthetics and design principles to ensure the visualizations are not only informative but also visually appealing.

The ultimate goal was to create clear, concise, and engaging visualizations that effectively communicate the insights gleaned from the data.

5.10 Building a Data Story in Tableau: A User-Friendly Journey

The chosen visualizations are not presented in isolation. Tableau allows us to create dashboards that weave these visualizations into a cohesive story. This story guides users through the key findings and insights revealed by the data analysis. Here's what this entails:

- **Logical Flow:** Visualizations are arranged in a logical sequence, guiding users from an overall picture to more granular details.
- **Interactive Elements:** Interactive features like filters and tooltips are incorporated to allow users to explore the data at their own pace and focus on specific areas of interest.
- **Annotations and Context:** Clear annotations and contextual information were added to provide background and explanation for the presented data.

By crafting a well-structured data story, SANTAR empowers users to engage with the analysis in a meaningful and interactive way. To test and use the dashboard please navigate to the link: https://github.com/Nishtha1263/SANTAR

5.11 Challenges Encountered: Lessons Learned

The journey of developing SANTAR wasn't without its hurdles:

- **Data Collection Time:** The sheer volume of historical data necessitated a significant investment of time during the collection phase. Utilizing more automated scraping techniques or leveraging existing datasets from reputable sources could potentially expedite this process in the future.

- **Ethical Scraping:** Balancing data acquisition needs with ethical considerations posed a challenge and alternative data sources should be explored if necessary.
- **Data Volume and Tableau:** Working with large datasets presented challenges in terms of smooth operation within Tableau. Restructuring the data or exploring cloud-based solutions for Tableau could mitigate these issues in future iterations.
- **Time Constraints:** Limited time impacted the scope of the project and the number of visualizations that could be created. Prioritizing key insights and focusing on a core set of visualizations can help manage time limitations.

5.12 Limitations Acknowledged: Areas for Improvement

While SANTAR provides valuable insights, there are limitations to consider:

- **Time Constraints:** The project timeline limited the depth and breadth of data exploration and visualization. With more time, additional data points (e.g., biomechanical data) could be incorporated, and a wider range of visualizations could be explored.
- **Data Availability:** The visualizations are based on the data that was successfully collected and cleaned. There may be limitations due to null values or errors in the data. Implementing more robust data cleaning techniques or exploring alternative data sources could potentially address these limitations.
- **Detailed Data:** The data scrapped is detailed but more insight can be provided if additional fields like split time, turn time, etc. are available.

5.13 Future Directions: Expanding the Horizons of SANTAR

Looking beyond the initial implementation, SANTAR presents possibilities for future development:

- **Data Expansion:**
  - Include Olympic data alongside World Championships to provide a more comprehensive picture of elite performance.
  - Broaden the scope to encompass regional and national championships to capture talent development pipelines.
  - Integrate biomechanical data like stroke rate, stroke length, and underwater dolphin kick duration to gain deeper insights into the technique.
- **Advanced Analysis:**

  - Utilize machine learning for performance prediction, training strategy optimization, competition country relevance, and even potential injury detection.
  - Conduct comparative analysis across countries, training philosophies, and swimming eras to identify best practices and emerging trends.
  - Develop functionalities for individual athlete performance monitoring, allowing coaches to tailor training plans more effectively.

- **User Experience Enhancements:**
  - Create a mobile app for on-the-go access to insights and visualizations for coaches and athletes.
  - Develop interactive storytelling features to engage users and facilitate a deeper understanding of the data.

6 Data Analysis

In the provided section, an in-depth analysis of various visualizations will be conducted. Each visualization will be scrutinized based on several factors, including the rationale behind selecting a particular type of graph, the insights that can be derived from the visualization, and the potential use cases for it.

These visualizations will be seamlessly integrated into the tableau story flow. For illustrative purposes within the report, the analysis will focus on the "50m Backstroke Women's" event as a reference point.

6.1 Country Performance:

- **Visualization Type:** World Map with Color Gradient and Marker
- **Description:** This world map employs a color gradient to represent a chosen metric (e.g., average rank, number of participants, average time, or average points) for a specific event across different countries. By hovering over the countries, users can see a tooltip highlighting the country's performance based on the average time for that event, the number of participants, and the average points.
- **Inference and Importance:** This visualization provides a quick and clear geographical snapshot of how different countries fare in a particular event. It can reveal the historical dominance of certain countries, prompting further investigation into training methods or cultural factors that might contribute to success.
- **Use Case:** Identify countries with consistently strong performances, potentially inspiring training strategies, or collaboration opportunities.

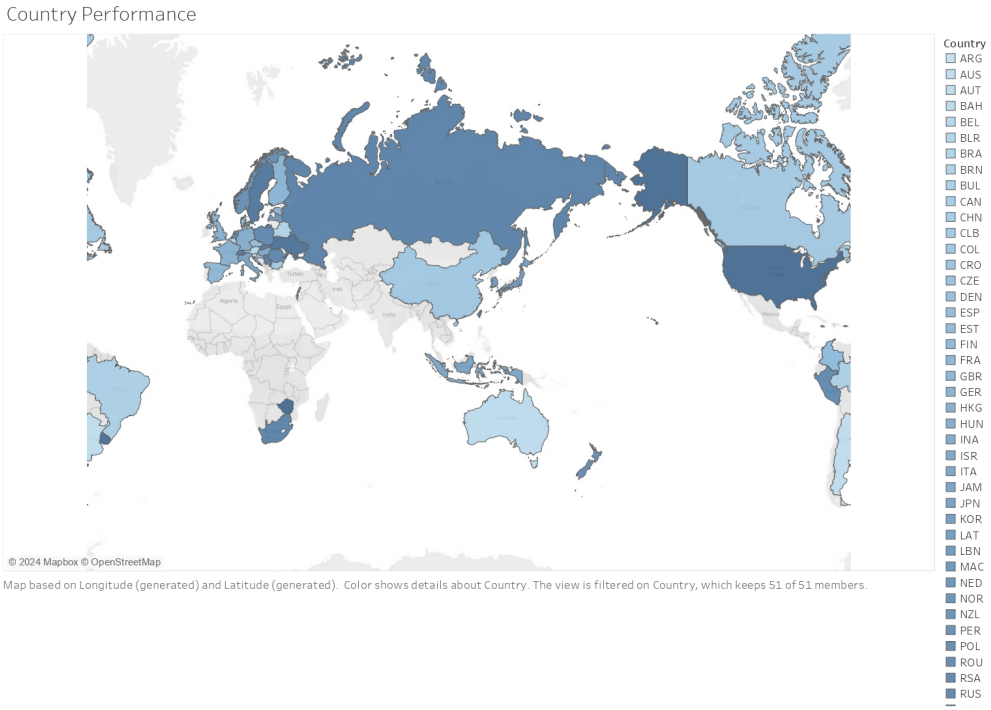

Figure 3: This world map utilizes a color gradient to depict the metrics with a tooltip (e.g., average rank, number of participants, etc.) for a given event across various countries.

6.2 Heats vs. Finals

- **Visualization Type:** Line Graph
- **Description:** This line graph tracks the performance (time) of a single swimmer in both the heats and finals of an event across multiple World Championships. Years with missing data might result in breaks in the line. There are also filters in this visualization to select the months and the swimmer to be analyzed.
- **Inference and Importance:** Swimmers often strategically pace themselves during heats, aiming to qualify for the finals rather than achieving their peak performance. This visualization allows coaches and athletes to analyze how a swimmer's performance progresses from heats to finals, revealing their pacing strategy and potential for improvement.

- **Use Case:** Evaluate a swimmer's ability to perform under pressure in the finals compared to the heats, informing training plans focused on maintaining peak performance throughout a competition.

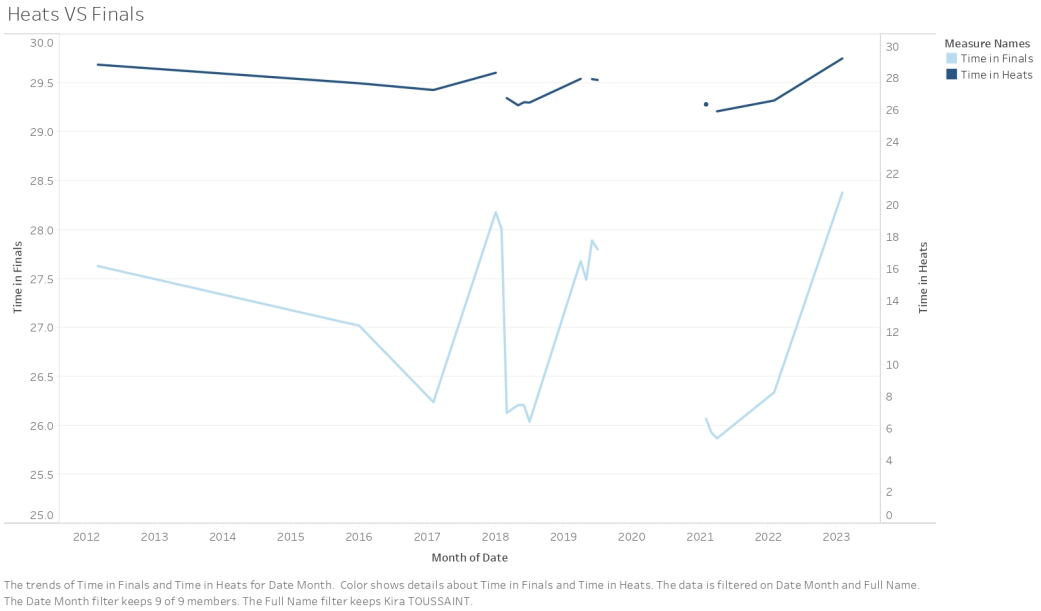

Figure 4: The Heats VS Finals time of a swimmer over the years in each World Championship stop.

6.3 Individual Performance:

- **Visualization Type:** Bar Graph
- **Description:** This bar graph depicts an individual swimmer's performance across multiple World Championships for a chosen event. It represents the metrics reaction time, rank, and time achieved in different colors. Filters provided allow the user to select a specific swimmer and the completion stop.
- **Inference and Importance:** This visualization provides a clear overview of an athlete's performance trajectory over time. It helps identify trends such as improvement,

stagnation, or decline, allowing coaches to tailor training programs accordingly.

- **Use Case:** Monitor an athlete's progress, set performance goals, and identify areas for improvement based on past performance trends.

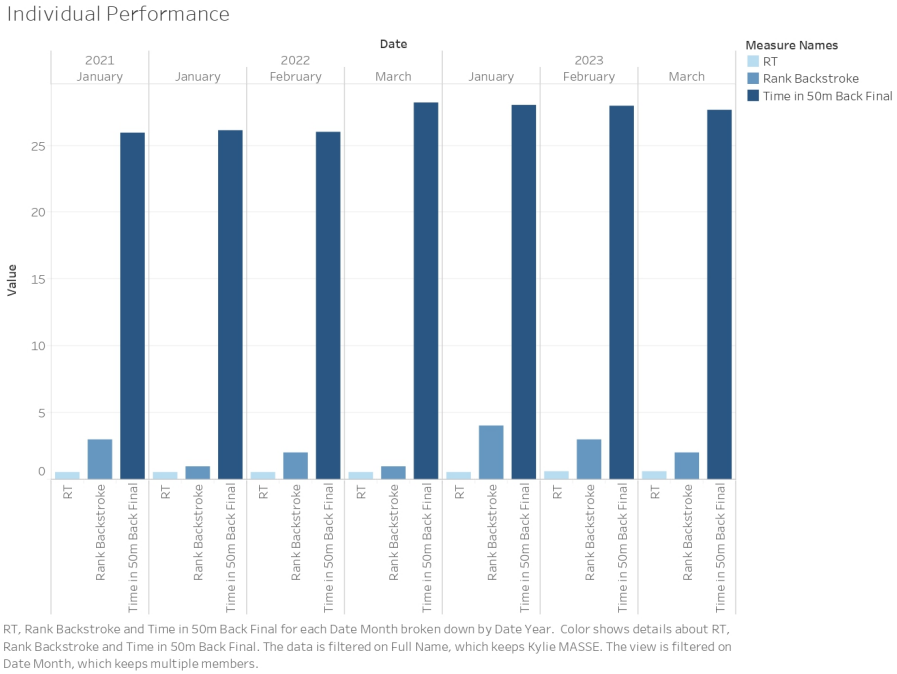

Figure 5: This bar graph illustrates a swimmer's performance in various World Championships for a selected event, showcasing reaction time, rank, and achieved time using distinct colors.

6.4 Same Stroke Different Distance:

- **Visualization Type:** Area Chart
- **Description:** This area chart showcases an athlete's performance (time) over multiple World Championships for different distances within the same stroke (e.g., freestyle across 50m, 100m, and 200m events). A filter is provided to select which swimmer's analysis the user wants and the legend tells the user what each color of the area chart is.

- **Inference and Importance:** This visualization reveals an athlete's strengths and weaknesses across different distances within a single stroke. It can identify if an athlete excels in shorter bursts or longer endurance swims, informing training strategies.

- **Use Case:** Develop training plans that target specific distances based on an athlete's performance profile.

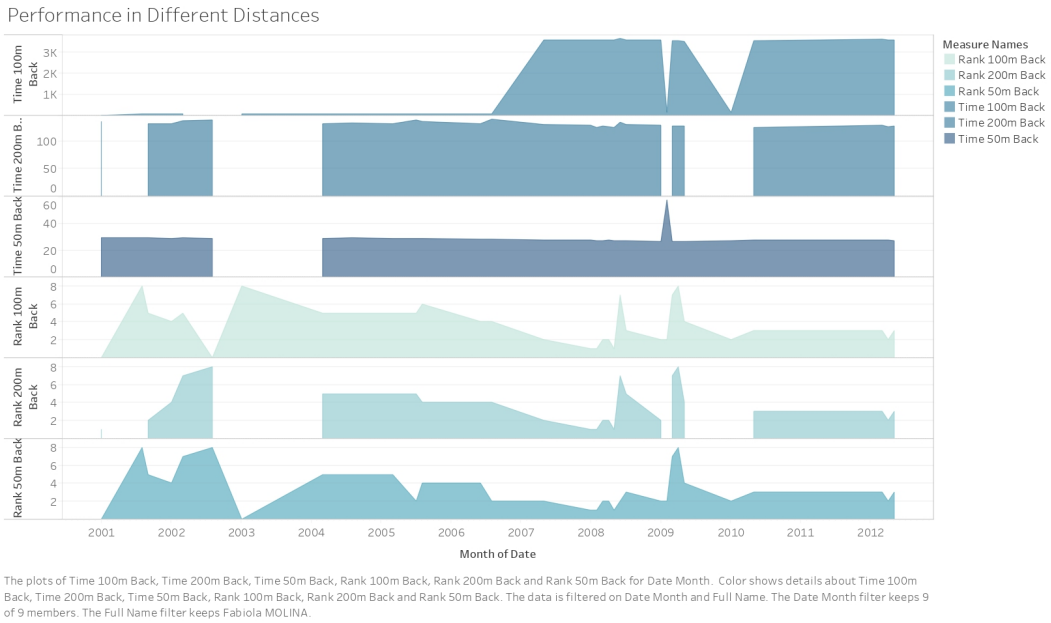

Figure 6: The area chart illustrates an athlete's performance (time) across multiple World Championships for different distances within the same stroke (e.g., freestyle for 50m, 100m, and 200m events).

6.5 Different Strokes Performance:

- **Visualization Type:** Side-by-Side Circle Chart

- **Description:** This visualization employs different color circles placed side-by-side. Each circle represents an individual swimmer's performance (average time or another

chosen metric) across different strokes (e.g., freestyle, butterfly, backstroke, breaststroke). The legend allows the user to see what each color circle is. There is also a filter given for the user to choose a swimmer of their preference.

- **Inference and Importance:** This visualization provides a quick comparison of an athlete's performance across different strokes. It reveals their strongest and weakest strokes, allowing coaches to identify areas for improvement and potentially develop training plans to achieve a more balanced skillset.

- **Use Case:** Evaluate an athlete's strengths and weaknesses across different strokes, guiding the development of a well-rounded training regimen.

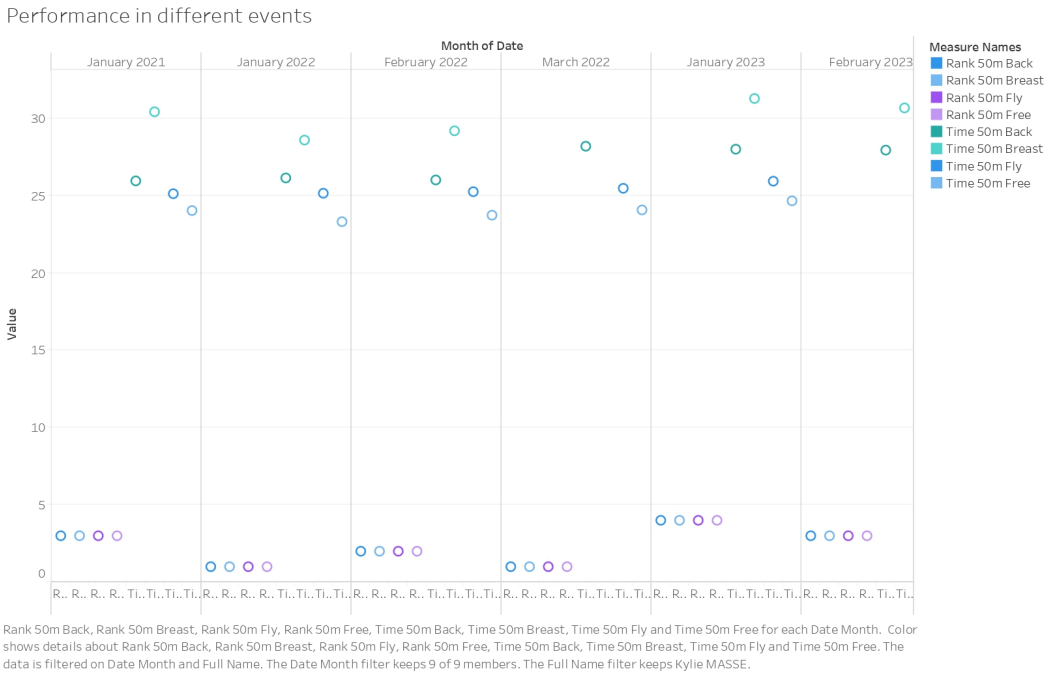

Figure 7: This visualization utilizes side-by-side colored circles, each representing an individual swimmer's performance (average time or another chosen metric) across various strokes (e.g., freestyle, butterfly,etc.).

6.6 Swimmer Comparison:

- **Visualization Type:** Side-by-Side Bar Graph
- **Description:** This visualization utilizes side-by-side bar graphs to compare the performance (time or another chosen metric) of two swimmers in a specific event across multiple World Championships. There are three different bars for each competition representing the three different metrics the swimmers are evaluated on.
- **Inference and Importance:** This visualization allows coaches and athletes to benchmark performance against competitors. It can reveal performance gaps, identify areas where one swimmer excels, and inform strategies to close the gap.
- **Use Case:** Analyze an athlete's performance relative to competitors, identify areas for improvement, and develop training plans to achieve a competitive edge.

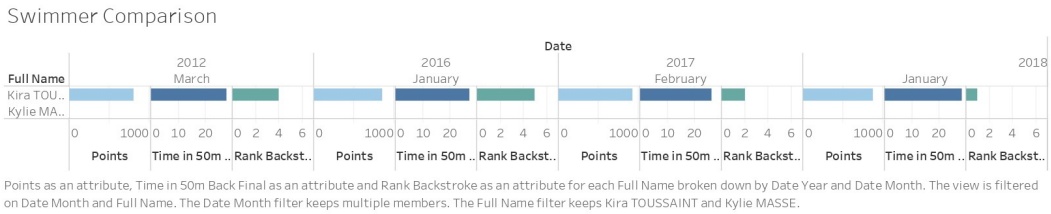

Figure 8: This visualization compares two swimmers' performance in a specific event across multiple World Championships using side-by-side bar graphs. Each competition features three bars representing different evaluation metrics.

6.7 Heat Map of Age vs. Time:

- **Visualization Type:** Heat Map
- **Description:** This heat map depicts the relationship between swimmer age and performance time for a specific event. Each cell in the heat map represents a combination of age and time.
- **Inference and Importance:** This visualization reveals trends in performance based on swimmer age. It can identify age groups with higher concentrations of top performers, potentially suggesting optimal training approaches or highlighting physiological factors influencing performance at different stages of an athlete's career. The graph also shows the decline in the number of swimmers between the ages of 15-20.

- **Use Case:** Evaluate age-related performance trends, identify potential peak performance windows, and tailor training plans to optimize performance throughout an athlete's career.

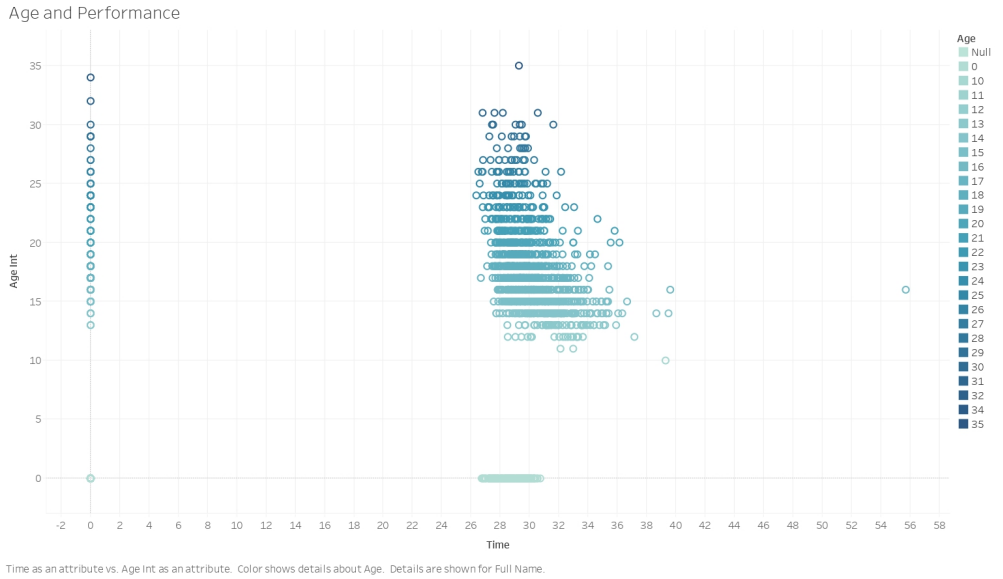

Figure 9: This visualization reveals trends in performance based on swimmer age by marking a swimmer's performance at a particular age.

6.8 Circle Chart of Outside Lane Wins and Losses:

- **Visualization Type:** Circle Chart
- **Additional Information:** Outside lanes in my visualization include lanes 1, 7, 8, and 9, and inside lanes include lanes 2, 3, 4, 5, and 6.
- **Description:** This circle chart is divided into sections, with each section representing a year for a specific event. Each year marks the number of outside lane wins and losses in two different colors.

- **Inference and Importance:** This visualization explores the potential influence of the starting lane on race outcome. Lanes closer to the center (lanes 4 and 5) might offer a slight confidence advantage, while outer lanes might present a psychological challenge. Observing win/loss distribution across the years can provide insights into these potential factors.
- **Use Case:** Analyze the potential influence of the starting lane on race outcome and develop strategies to mitigate any disadvantages associated with outer lanes.

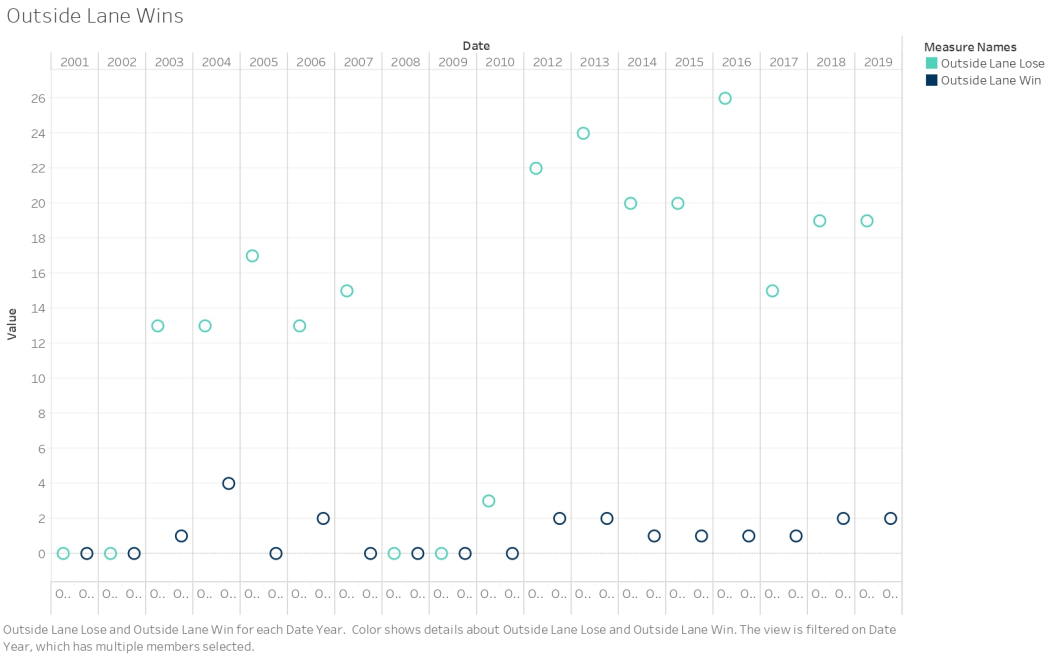

Figure 10: Outside Lane Wins and losses in a year marked din two different colors.

7 Conclusion

This report chronicles the journey of SANTAR, a project dedicated to exploring the fascinating world of competitive swimming through the lens of data analysis. By meticulously collecting and cleaning data from World Championships, SANTAR paints a detailed picture

of performance trends and unveils insights that can empower coaches, athletes, and swimming enthusiasts alike.

The initial focus of SANTAR was on gathering data from the World Championships, a pivotal event for swimmers, serving as both a stepping stone to the Olympics and a platform for strategic experimentation. The years 2000-2023 were chosen to provide a comprehensive historical snapshot. A well-organized data structure was established, with dedicated folders for each year and separate workbooks categorizing results by specific competitions within each year. This facilitated efficient data storage and retrieval.

Utilizing Python libraries like `selenium` and `BeautifulSoup`, data was extracted from the official FINA website. Ethical considerations were paramount, adhering to guidelines and respecting data licensing terms. Once collected, the raw data underwent a cleaning process. Inconsistencies in formatting were addressed and missing values were handled. These steps ensured the data's accuracy and usability for subsequent analysis.

Tableau, a powerful data visualization software, was employed to transform the cleaned data into clear and informative visuals. Moving beyond raw numbers, these visualizations communicate trends and patterns in an accessible way, allowing users to grasp complex information without extensive statistical expertise.

SANTAR didn't simply present a collection of isolated visualizations. Tableau's functionalities allowed for the creation of a cohesive data story. Visualizations were arranged in a logical flow, guiding users from an overall picture to more granular details. Interactive elements like filters and tooltips enhanced user engagement, allowing them to explore the data at their own pace. Contextual information and clear annotations further enriched the story, providing background and explanation for the presented findings.

SANTAR utilizes eight compelling visualizations, each offering a unique perspective on swimming performance:

1. **Country Performance (World Map):** The visualization reveals the historical dominance of specific countries in various events. By employing a color gradient and markers, it provides a global snapshot of performance metrics (average rank, number of participants, average time) across different countries.

2. **Heats vs. Finals (Line Graph):** The line graph charts the performance of a single swimmer in both heats and finals across multiple World Championships. It sheds light on the "save for finals" strategy, where swimmers might pace themselves strategically during heats to qualify for the finals.

3. **Individual Performance (Bar Graph):** Tracking an athlete's progress over time, this visualization depicts metrics like reaction time, rank, or time achieved in an event across multiple World Championships. It helps identify trends in performance, aiding coaches in tailoring training programs.

4. **Same Stroke Different Distance (Area Chart):** The visualization reveals an athlete's strengths and weaknesses across different distances within the same stroke

(e.g., freestyle across 50m, 100m, and 200m). Analyzing the area under the curve for each distance informs training strategies focused on specific distances.

The remaining visualizations delve even deeper:

5. **Different Strokes Performance (Side-by-Side Circle Chart):** Comparing an athlete's performance across different strokes (freestyle, butterfly, backstroke, breaststroke), this visualization helps identify areas for improvement and enables the development of a well-rounded training regimen.

6. **Swimmer Comparison (Side-by-Side Bar Graph):** Benchmarking performance against competitors, this visualization allows coaches and athletes to identify areas where one swimmer excels, informing strategies to close the performance gap.

7. **Heat Map of Age vs. Time (Event):** This heat map unveils age-related trends. Observing the frequency of swimmers achieving specific times at different ages provides insights into potential peak performance windows and informs training approaches tailored to different career stages.

8. **Circle Chart of Lane Wins/Losses (Event):** Analyzing the potential influence of the starting lane on race outcome, this visualization examines the distribution of wins and losses across different lanes. This can help develop strategies to mitigate any disadvantages associated with outer lanes.

SANTAR presents a springboard for further exploration. Expanding the data scope to include Olympic data and regional/national championships offers a more comprehensive picture of talent development. Integrating biomechanical data like stroke rate and underwater dolphin kick duration delves into the realm of technique analysis. Machine learning holds immense potential for performance prediction, training strategy optimization, and even injury detection. Imagine a future where SANTAR empowers coaches to personalize training plans based on an athlete's strengths, weaknesses, and predicted performance potential. The possibilities for performance enhancement are truly exciting.

The journey of developing SANTAR wasn't without its hurdles. The sheer volume of historical data necessitated significant time investment during collection. Exploring alternative data sources or leveraging more automated scraping techniques could expedite this process in the future. Ensuring ethical scraping practices and respecting data licensing terms remained a crucial consideration. Additionally, working with large datasets presented challenges within Tableau, highlighting the need for potential solutions like data restructuring or exploring cloud-based alternatives. Time constraints inevitably limited the scope of the project and the number of visualizations that could be created. However, prioritizing key insights and focusing on a core set of visualizations mitigated this limitation.

SANTAR has taken a significant step towards demystifying the world of competitive swimming. By harnessing the power of data analysis and visualization, SANTAR offers

valuable insights that can empower various stakeholders. Coaches can leverage these insights to tailor training plans, identify areas for improvement, and benchmark performance against competitors. Athletes gain a deeper understanding of their own strengths and weaknesses, track their progress over time, and set realistic goals. Swimming enthusiasts can appreciate the rich tapestry of factors that contribute to achieving excellence in the pool.

Looking forward, SANTAR presents a roadmap for continuous exploration. By expanding the data scope, incorporating biomechanical data, and delving into advanced analysis techniques, SANTAR holds the potential to revolutionize our understanding of swimming performance. Ultimately, SANTAR aspires to be a valuable tool that propels swimmers toward achieving their full potential and reaching the pinnacle of aquatic excellence.

8 Acknowledgment

The development of SANTAR would not have been possible without the invaluable support and guidance of several individuals. I would like to express my sincere gratitude to my faculty mentor, Mr. Chiranjoy Chattopadhyay, Professor of Computer Science at FLAME University, Pune. Mr. Chiranjoy Chattopadhyay's expertise in this field and unwavering support were instrumental in shaping the direction of this project. Their insightful feedback and encouragement throughout the research process were crucial to my success.

I would also like to extend my appreciation to the FLAME University community for providing the resources and environment necessary for this project to flourish. The access to the computing facilities proved invaluable in conducting the data analysis and visualization tasks. I would also like to thank the university for providing me the opportunity to work on this project.

Finally, a thank you to my colleagues and peers who offered their time and feedback during the development of SANTAR. Their constructive criticism and willingness to engage in discussions helped refine my approach and ensure the clarity of the final product.

References

[1] N. Smith and J. Norris, *Analysis of Swimming Performance and Technique [Chapter 13]*, N. Armstrong and D. Mitchell, Eds. Churchill Livingstone, 2002.

[2] B. T. M. Morais, H. P. and A. M. Costa, "A study of pacing strategies in fina world championships swimming finals," *International Journal of Performance Analysis in Sport*, vol. 20, no. 3, pp. 828–840, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/13/18/10515

[3] W. Z. W. Y. G. H. Liu, Z. and Z. Yu, "The relationships between anthropometric characteristics and competitive swimming performance of elite athletes," pp. 592–597, 2013. [Online]. Available: https://dialnet.unirioja.es/descarga/articulo/7950307.pdf

[4] H. M. Toussaint and P. J. Beek, "Relationship between underwater dolphin kicking duration and front crawl performance," pp. 147–154, 1992. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33593229/

[5] "World aquatics." [Online]. Available: https://www.worldaquatics.com/

[6] Tableau Software, "Tableau creator: Get started." [Online]. Available: https://www.tableau.com/learn/get-started/creator