Visualization Project Report

Multi-Year Road Race Analysis: Participation and Performance Insights (2021-2025)

DASC 4231 | Final Project

1. Introduction and project objective

This project analyzes five years of road race data (2021–2025) to better understand participation trends, athlete demographics, and performance differences across race distances. Using Tableau, an interactive dashboard was developed to explore how participation has changed over time, how different groups of runners are represented, and how performance varies by age, gender, and race type.

The goal of this project is to provide meaningful insights that can help race organizers improve event planning, increase participation, and better target different groups of runners. By identifying trends in participation, performance, and geographic distribution, this analysis supports more data-driven decision-making for future races.

2. Dataset description

The dataset used in this project consists of five Excel files containing road race registration and results data from 2021 to 2025. These files were combined into a single dataset containing a total of 2,409 participant entries.

Table 1. Participant Entries by Source Year

Race year	Source file	Participant entries
2021	21-09-04.xls	328
2022	22-09-03.xls	407
2023	23-09-02.xls	407
2024	24-08-31.xls	630
2025	25-08-30.xls	637
Total	**All files**	**2,409**

The data includes three race distances: 1K, 5K, and 10-mile events. It contains 34 features covering race details, participant demographics, performance metrics, and participation status. Key variables include age, gender, race group, city, state, finishing times, pace, and placement rankings.

The geographic scope of the dataset is primarily Texas-based, with some participants from other states. The dataset also includes indicators for whether participants registered but did not start, allowing for analysis of participation and dropout trends.

The dataset required preprocessing to ensure consistency across years. Some fields were formatted differently between source files, particularly time-based variables, which required standardization into numeric formats. Additionally, minor data quality issues such as missing or incomplete values were addressed during cleaning. These steps were necessary to ensure accurate comparisons across years, race distances, and participant groups.

3. Methods and Tools Used

Data preparation and analysis were conducted using Python and Tableau. The raw Excel files were first merged into a single dataset using Google Colab. Several preprocessing steps were performed to clean and structure the data for analysis.

These steps included extracting the race year from file names, converting time variables into numeric formats, and splitting combined demographic fields into separate variables. Additional features were created, such as a “did not start” indicator and pacing difference metrics for longer races.

Additional preprocessing steps included standardizing time variables into comparable formats, handling missing values through filtering and cleaning, and grouping ages into defined bins to support demographic analysis. These transformations ensured that the dataset was structured consistently and suitable for both visualization and comparison across multiple years.

Tableau was then used to build an interactive dashboard consisting of multiple visualizations. Filters for race year and distance were applied across all charts to allow for dynamic exploration of the data.

4. Visualization Design and Rationale

The dashboard was organized around two main views: participation and performance. This separation keeps the analysis easier to follow. Participation uses the full dataset, while performance is limited to only the participants with valid, recorded times. Participation visuals answer questions about who registered, how many people participated, and where they came from. Performance visuals focus on finish times, pace, age, gender, and distance.

The design uses simple chart types because the audience is likely to include race organizers, volunteers, and other stakeholders who need quick answers rather than a highly technical display. Chart types were selected based on the nature of the data and the type of comparison required. Bar charts were used for categorical comparisons, such as participation by age group and gender, because they allow for clear side-by-side evaluation. Line-based views were used to highlight trends over time, such as changes in participation across years. Distribution plots were included to show variability in finish times, which provides more insight than summary statistics alone. These choices were made to ensure the dashboard is both intuitive and effective for non-technical stakeholders.

The dashboards also use filters and color groupings to make exploration more natural. For example, users can filter by race, year, or distance, then compare whether participation and performance patterns change. This interactive structure is more useful than a static report because the same dashboard can answer several related questions without requiring a new chart each time.

Table 2. Dashboard Visualizations and Design Purpose

Dashboard element	**What it shows**	**Why was it used**
Race KPI Summary	Total registrants, DNS rate, average pace, and year-over-year growth by year and distance.	Provides a quick summary before the user views the detailed charts.
Participation by Year and Distance	How participation changes across years and race distances.	Makes the growth pattern and distance comparison easy to see.
Age Group Participation by Gender	Participation levels across age bins and gender.	Highlights which demographic groups are most represented and which may need more outreach.
Gender and Distance Participation	Male and female participation across race, distance, and year.	Shows whether participation patterns differ by gender and race type.
Geographic Distribution Map	Participant locations by city and state, with size based on runner count.	Helps identify local concentration and potential areas for regional marketing.

Performance Analysis	Average chip time or pace by age group and gender.	Connects demographic groups with performance patterns.
Finish-Time Distributions	Spread of finish times by distance, age group, and gender.	Shows variability rather than only showing averages.
Top Finishing Times	Fastest times by year, gender, and distance.	Provides a competitive-performance view for the strongest finishers.

5. Key Findings

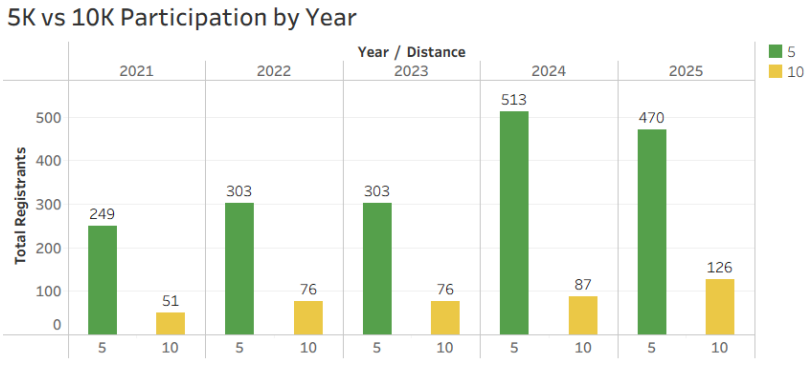
The analysis of the multi-year race dataset reveals several meaningful trends in participation, demographics, and performance that can support more informed decision-making for future events.

Participation trends showed strong overall growth, with a major increase in 2024.

Participation remained relatively stable between 2022 and 2023, followed by a significant surge in 2024. Registrations increased by over 200 participants in a single year, representing the largest year-over-year growth in the dataset. Although participation slightly declined in 2025, it remained well above that of earlier years, indicating continued interest in the event. This pattern may reflect increased event visibility, improved marketing efforts, or continued post-pandemic recovery in group activities.

Mid-distance races attracted the highest participation.

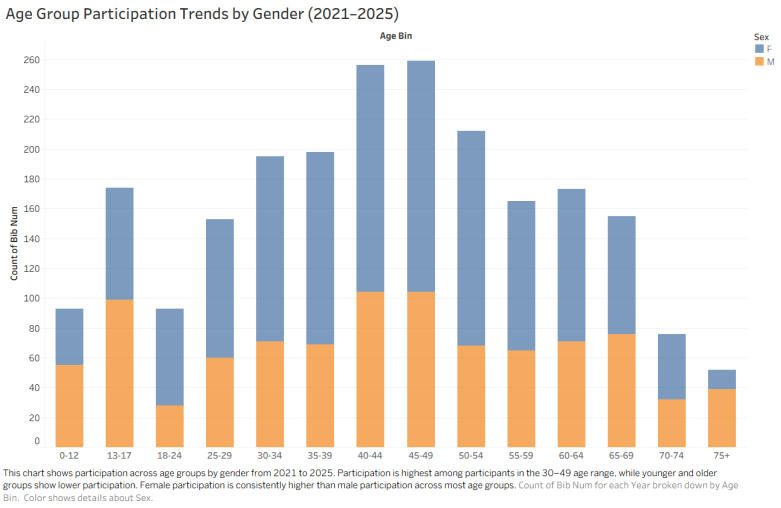
The *Participation by Year and Distance* chart shows that the 5K distance consistently has the highest number of participants across all years. In addition, female participation appears especially strong in the 5K across multiple views. This suggests the 5K is the most accessible and widely appealing race option, particularly among female runners.



This chart compares participation in the 5K and 10K races across multiple years. Participation has increased over time for both distances, with the 5K consistently attracting significantly more runners than the 10K race. Count of Bib Num for each Distance broken down by Year. Color shows details about Distance. The data is filtered on Race Year (participation_df), which keeps 21, 22, 23, 24 and 25.

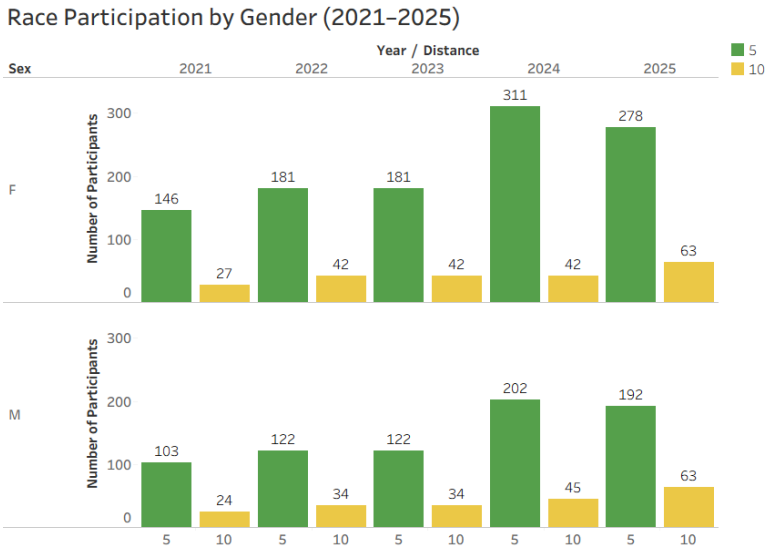
Participation was concentrated in specific age groups.

The *Age Group Participation by Gender* chart shows that the majority of participants are in the 20–50 age range, with lower representation in younger and older groups. This suggests opportunities to expand outreach to underrepresented age groups and broaden overall participation.



Gender participation was relatively balanced but varied by distance.

The *Gender and Distance Participation* chart shows that male and female participation is generally balanced overall, although female participation is slightly higher in several views, particularly in the 5K. Differences across race distances suggest gender-based preferences and highlight where engagement is strongest.



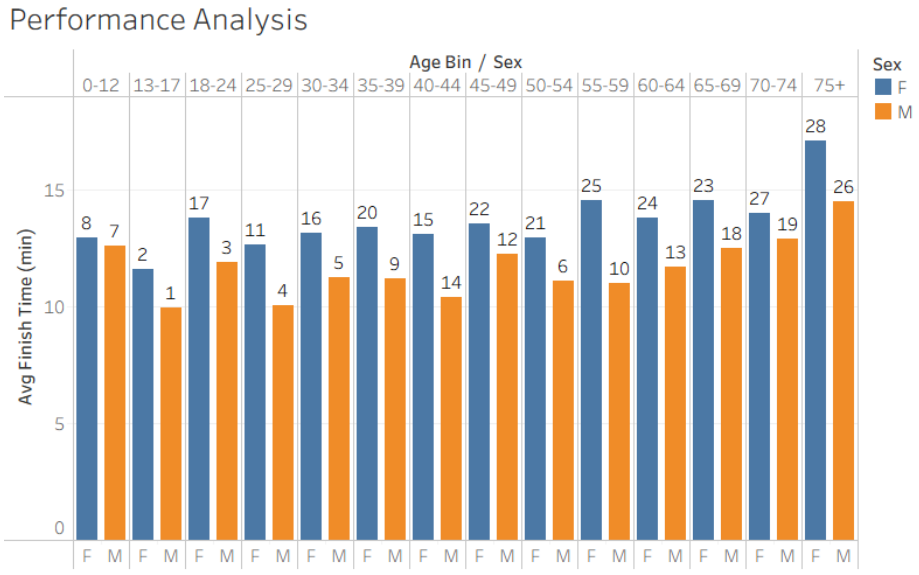
Geographic participation was primarily local.

The *Geographic Distribution Map* shows a high concentration of participants from Texas, particularly near the race location, with fewer participants from out-of-state regions. This indicates that the event is largely driven by local engagement and may benefit from expanded regional outreach.

This map shows where race participants are coming from across the United States. Most participants are concentrated in specific regions, likely near the race location, with smaller numbers traveling from more distant areas. This information can help guide regional marketing efforts. Map based on Longitude (generated) and Latitude (generated). Color shows details about Year. Size shows sum of Runner count. Details are shown for City and State. The data is filtered on Race Year (participation_df), which keeps 21, 22, 23, 24 and 25.

Performance trends aligned with expected patterns.

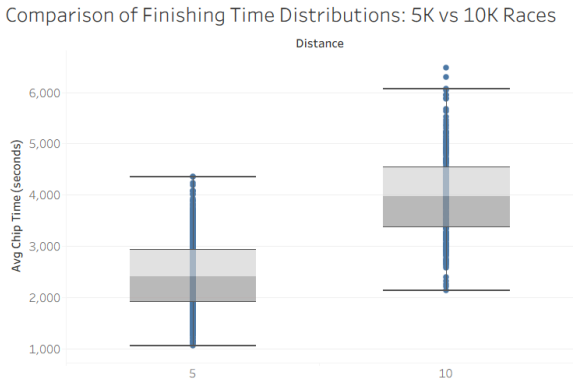
The *Performance Analysis* chart demonstrates that average pace and finish times vary by age group and gender, with younger participants generally achieving faster times. Finish times also tend to increase with age; however, noticeable variation within each age group suggests that factors beyond age, such as fitness level and experience, also influence performance.



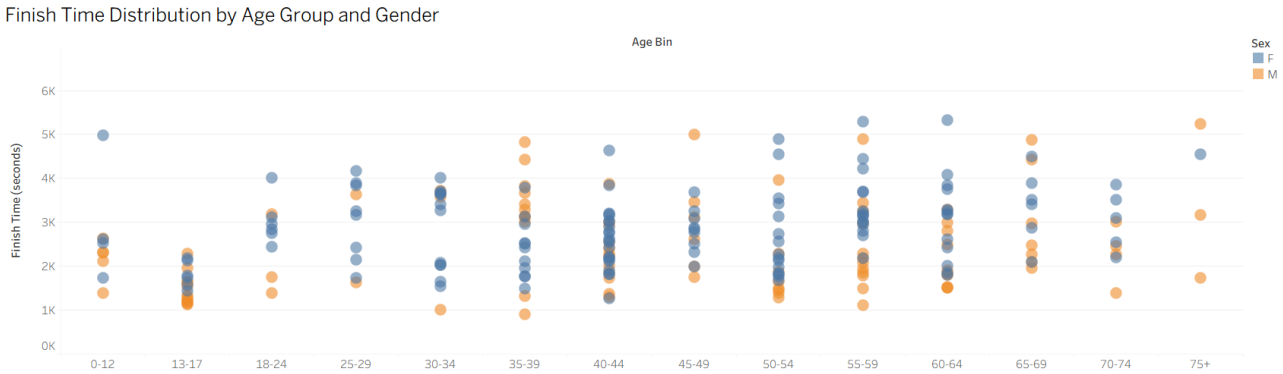
This chart compares average finish times across age groups and gender. Finish times generally increase with age, and male participants tend to have slightly faster times than female participants across most age groups. Selected Metric for each Sex broken down by Age Bin. Color shows details about Sex. The marks are labeled by Finish Rank. The data is filtered on Distance Label and Year. The Distance Label filter keeps 10-Mile and 5K. The Year filter keeps 2021, 2022, 2023, 2024 and 2025. The view is filtered on Sex, which keeps F and M.

Finish-time distributions revealed wide variability among participants.

The finish-time distribution charts show a broad range of completion times across race distances, particularly in the 10K races, reflecting differences in runner experience, pacing, and endurance levels. The wider spread and higher median times in longer races suggest greater performance variability compared to the 5K events. When analyzed by age group and gender, finish times also display noticeable variation within each category, though overall trends indicate that average completion times generally increase with age.



This chart compares the distribution of finishing times between the 5K and 10K races. The 10K race shows a wider spread and higher overall finish times, indicating greater variability and endurance requirements. In contrast, the 5K results are more tightly grouped, suggesting more consistent performance. Average of Chip Time Sec for each Distance. Details are shown for Distance and Bib Num. The data is filtered on Year, which keeps 2021, 2022, 2023, 2024 and 2025.



This visualization shows the distribution of finish times across age groups for both genders. There is noticeable variability within each group, but overall trends indicate increasing finish times with age. This chart highlights the range of performance levels within the participant pool. Chip Time Sec for each Age Bin. Color shows details about Sex. Details are shown for Bib Num. The data is filtered on Distance, which keeps multiple members. The view is filtered on Year, which keeps 2021, 2022, 2023, 2024 and 2025.

Participation gaps, such as DNS, impact overall event outcomes.

While not shown in a single chart, the dataset includes “Did Not Start” (DNS) participants, which directly affect race-day logistics and planning.

This suggests that improving communication, reminders, and small incentives (such as packet pickup engagement) could help reduce no-show rates and improve overall event efficiency.

These findings directly address the core questions for this project by demonstrating how participation evolved, how it differs across each race, distance, and demographic group, and where participants are geographically concentrated. The analysis also highlights the impact

non-starters have on overall event outcomes. Together, these insights provide clear answers to the questions the team set out to explore.

6. Limitations and Future Improvements

There are a few limitations to keep in mind. First, this project is based on one annual race series, so the findings should not be treated as a general rule for all running events. The results are most useful for understanding this specific race and its local participant base.

Second, finish time is influenced by more than age, gender, or distance. Weather, course conditions, training level, injuries, race-day delays, and whether someone ran or walked can all affect performance. Those factors were not fully captured in the dataset, so the performance results should be interpreted as patterns rather than explanations of why each runner finished at a certain time.

Third, the dataset includes both registration and performance-related records. That is useful, but it also means the analysis has to be careful when comparing registrants, starters, and finishers. Future versions should keep these categories clearly separated so that participation counts and performance summaries are not accidentally mixed.

Future improvements could include:

- Adding weather, course, and race-day condition data so performance changes can be interpreted more fairly.
- Tracking repeat participants to measure retention and identify who returns year after year.
- Collecting registration source or marketing campaign data to see which outreach channels actually bring runners in.
- Adding clearer distance-standardization rules so the race labels stay consistent across raw files, processed data, slides, and Tableau visuals.
- Publishing a cleaned version of the Tableau dashboard with a short user guide so organizers can reuse it after future races.
- Expanding accessibility checks by keeping captions, meaningful chart titles, readable labels, and consistent color use throughout the workbook.

While the dashboard is functional today, these improvements could be made into a reliable reporting tool. Each year, organizers could load the new results, compare them with prior years, and make better decisions about promotion, race-day staffing, and participant experience.

7. Conclusion

This project successfully analyzed five years of road race participation and performance data to identify meaningful trends across race distances, demographic groups, and participant outcomes. Using Python for preprocessing and Tableau for visualization, the project transformed raw race records into an interactive dashboard designed to support data-driven decision-making for race organizers and stakeholders.

The analysis revealed several important findings. Participation increased significantly in 2024 and remained strong in 2025, with the 5K race consistently attracting the highest number of participants. Demographic analysis showed that most runners were between the ages of 20 and 50, while geographic analysis demonstrated that participation was largely concentrated within Texas and nearby regions. Performance-focused visualizations also showed that finish times generally increased with age and varied considerably across race distances, genders, and participant groups.

Although the project has some limitations, including the focus on a single race series and the absence of external performance factors such as weather or training conditions, the dashboard still provides a strong foundation for future analysis and reporting. Overall, this project demonstrates how interactive visualizations can transform race data into meaningful insights that support better event planning, participant engagement, and data-driven decision-making for future races.

Resources

Purple Monkey Fun Run: https://purplemonkeyfunrun.com/

Campbell Timing Systems: https://campbelltimingsystems.com/