**Marathon Results - 2017**

**Introduction:**

In this report, we delve into a detailed analysis and visualization of the Boston Marathon finishers' data from the 2017 event. Our objective is to uncover significant trends, demographic patterns, and performance metrics within the dataset. We aim to highlight key aspects, such as the distribution of finishers by age, providing insights into the prevalent age demographics among marathon participants. Moreover, we will examine participation rates across different countries, revealing the global reach and geographic diversity of the event. Additionally, our analysis will focus on the age distribution by gender, exploring potential variations and disparities in participation and performance based on gender identity. This comprehensive exploration will offer a nuanced understanding of the marathon's demographic and geographic landscape.

**Analysis and Visualization Approach:**

For this project, we planned to create three primary visualizations to explore the Boston Marathon finishers' data from 2017.

1. The first was a **Histogram** to illustrate the distribution of finish times, allowing for dynamic exploration of how runners performed.
2. The second was a **world map** displaying the geographical origins of participants, highlighting the global nature of the event.

1. The third was a **linear model visualization** to examine how age and gender influenced finish times, represented by the model’s coefficients.

**Data Preparation and Cleaning:**

To prepare the data for these visualizations, we undertook several key steps:

1. **Loading and Inspecting the Data**:

We began by loading the marathon\_results\_2017.csv file and examining its structure to understand the types and distributions of the data it contained.

1. **Data Cleaning**:

This included converting the Official Time from a string format into a numeric value representing total minutes. We also ensured all necessary fields (like age and gender) were appropriately formatted for analysis.

1. **Handling Missing Data**:

We checked for any missing or inconsistent entries, particularly focusing on the country field to ensure accurate geographical mapping.

1. **Merging with World Shapefiles**:

For the geographical visualization, we downloaded a world shapefile and merged it with the marathon data. Special attention was required to correctly map countries like Norway and France, which involved adjusting the country codes to align with the shapefile’s format.

1. **Feature Engineering**:

We created additional columns, such as Official\_Time\_Minutes, to facilitate the modeling and visualization processes.

**Visualizations:**

### **Distribution of Marathon Finish Times:**

### **A graph of a running Description automatically generated**

The histogram depicts the finish times of participants in the 2017 Boston Marathon. Most runners completed the marathon around the 4-hour mark, evidenced by the peak in the histogram. This indicates a common target for many amateur runners. The distribution is right-skewed, showing a gradual decline in the number of finishers beyond 4 hours and fewer participants finishing under 3 hours, highlighting the challenge of faster completion times. The long tail extending towards 7 hours reflects the diversity in runner performance, from competitive athletes to those achieving personal milestones. This visualization effectively captures the range and central tendency of finish times, providing insights into the endurance levels of the marathon participants.

1. **Marathon Participants by Country:**

**A map of the world

Description automatically generated**

The above visualization shows a world map indicating the number of marathon participants by country. The map uses a gradient color scale from light green to orange to represent participant numbers, with green indicating lower participation and orange indicating higher participation. The United States stands out with the highest participation, reflected by the deep orange shade, indicating around 20,000 participants. Several countries in Europe, Asia, and South America are shaded in lighter green, showing moderate levels of participation. In contrast, many countries in Africa, Central Asia, and parts of South America are light gray, indicating no data or negligible participation. The color gradient bar at the bottom of the map provides a reference for the participant numbers corresponding to the color intensity.

1. **Linear model predicting finish time based on Age and Gender:**

**A graph with blue and black text

Description automatically generated with medium confidence**

The visualization shows the coefficients of a linear regression model predicting marathon completion times (in minutes) based on participants' age and gender. The intercept coefficient, which represents the baseline predicted time for a female participant with an age of zero, is approximately 206.62 minutes. The age coefficient is around 1.06, indicating that each additional year of age increases the expected marathon time by about 1.06 minutes. The gender coefficient, labeled as 'M/F' M, is -25.30, suggesting that male participants generally complete the marathon approximately 25.30 minutes faster than female participants, holding age constant. The error bars indicate the standard errors of the coefficients, with all coefficients being highly statistically significant (p < 2e-16). The residual standard error is 39.18, and the model explains about 13.61% of the variability in marathon times, as indicated by the R-squared value.

1. **Linear Modelling Prediction and Further Analysis:**

A collage of graphs and charts

Description automatically generated

A graph showing a number of blue and pink dots

Description automatically generated

The scatter plot shows the relationship between the official finishing times (in minutes) and the ages of marathon participants, which distinguishes the gender of male and female runners by color. Blue dots represent male finishers, while pink dots represent female finishers. The black dashed line illustrates the linear regression model’s predicted finishing times based on age and gender. The plot reveals that finishing times tend to increase with age, and on average, males generally have faster finishing times compared to females. The spread of data indicates variability in finishing times across different ages and genders, highlighting the complex factors influencing marathon performance.

**Challenges and Future Approaches:**

Creating these visualizations posed several challenges. Converting and accurately plotting finish times required meticulous handling to prevent errors. Mapping participants' countries was complicated by the need to resolve discrepancies in country codes, particularly for Norway and France. Despite these challenges, the resulting visualizations offer valuable insights. Future exploration could involve segmenting data by more specific demographics, analyzing trends over multiple years, or incorporating additional variables like training intensity or race-day weather conditions to enhance predictive modeling.

**Application of Data Visualization Principles:**

For this assignment, the principles of data visualization and design were applied in several keyways:

1. **Clarity and Simplicity**: The visualizations were designed to convey information clearly and concisely. By using color gradients and bar charts, complex data was made easily interpretable.
2. **Appropriate Use of Color**: Color gradients were used effectively to represent different ranges of marathon participants, with a clear legend to interpret the colors. This helped in visually distinguishing between different levels of participation.
3. **Accuracy and Precision**: Care was taken to ensure that data conversions and plotting were done accurately to avoid misleading representations. This included correctly mapping finish times and resolving country code discrepancies.
4. **Statistical Significance**: The linear model coefficients were presented with error bars to show the precision of estimates, emphasizing the statistical significance of the findings.
5. **Informative Legends and Labels**: Each visualization included clear titles, legends, and labels. This helped in guiding the viewer to understand the context and scale of the data presented.
6. **Insights and Interpretability**: The visualizations were designed not just to display data, but to provide valuable insights. For example, the map of marathon participants highlights geographical trends, while the linear model coefficients chart reveals the impact of age and gender on finish times.
7. **Future Directions**: Suggestions for further analysis were included, such as segmenting data by more specific demographics or incorporating additional variables. This approach not only presented the current findings but also opened pathways for deeper exploration.

These principles ensured that the visualizations were both effective in communicating the data and insightful for further analysis.

**Conclusion:**

In conclusion, the creation of these visualizations successfully highlighted key insights into marathon participation and performance. Despite challenges in data handling, such as converting finish times accurately and resolving country code discrepancies, the visualizations effectively conveyed meaningful patterns. The map showcased geographical trends in marathon participation, while the linear model coefficients provided a clear understanding of the impact of age and gender on finish times. Future analyses could enhance these findings by incorporating additional variables like training intensity or race-day weather conditions, and by segmenting the data more specifically. Overall, this assignment demonstrates the power of well-designed visualizations in revealing significant trends and guiding further research.