Data Visualization Final Project

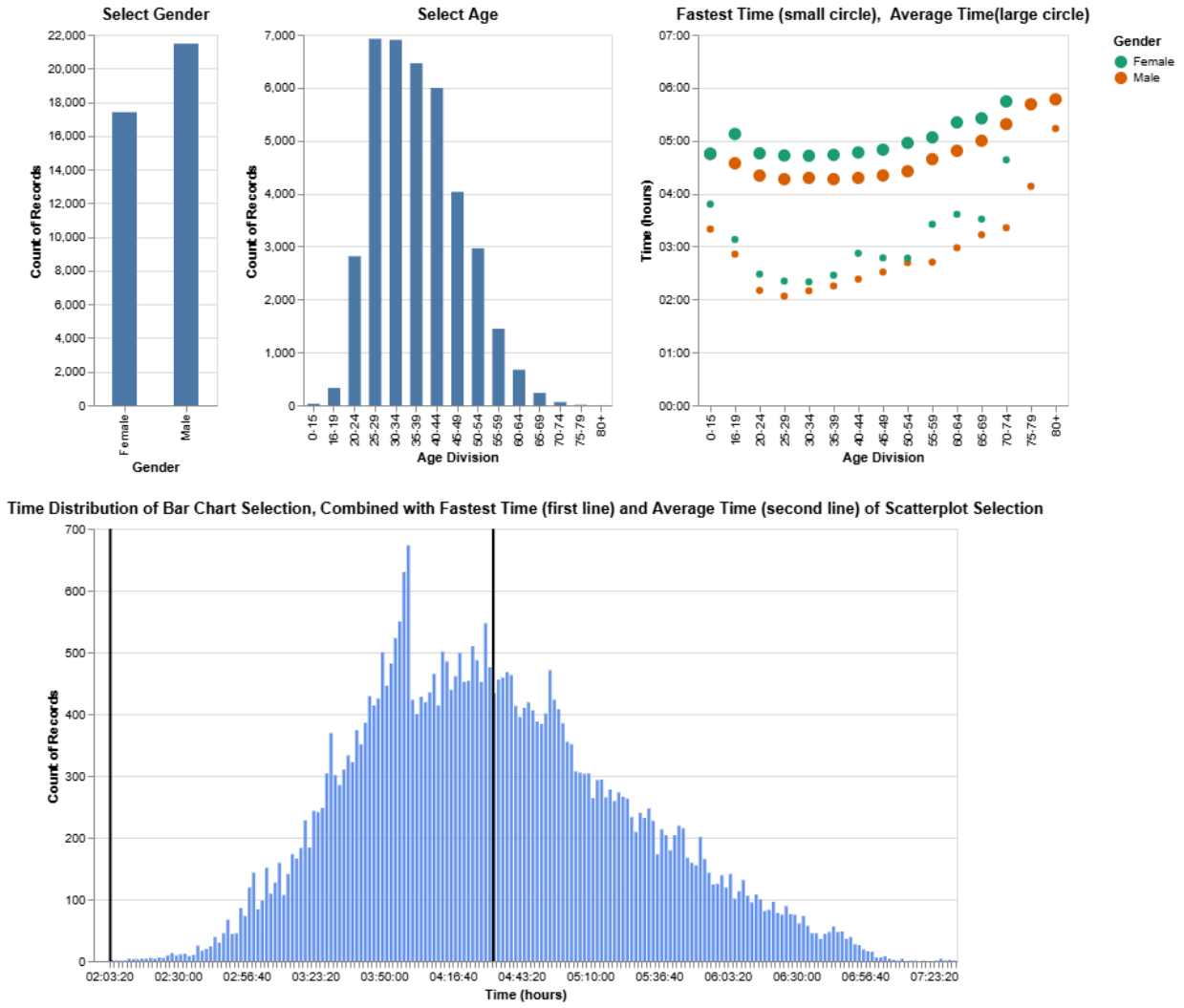
Introduction

For my final project I chose to create a visualization tool for exploring the results of the Chicago Marathon. The data for this project came from a public GitHub repository (https://github.com/mallaham/bofa-chicago-marathon) containing a file with results from every Chicago Marathon from 1996 to 2014. I chose to focus on the 2013 race for my project, as it was the most recent year with a complete dataset. After some data cleaning and preprocessing, I obtained a data file containing information about each of the nearly 40,000 runners who finished the race, including their name, country of residence, gender, age (within a 5 year interval), and finish time.

My goal for this project was to create an interactive tool with which users could gain a detailed understanding of the population of runners involved in the marathon and how their results are distributed. One of the first tasks that comes to mind is to visualize the demographics of the people involved in terms of the number of people of each gender, age group, and country. Another important task is to show the overall distribution of finish times to give users an impression of the typical time it takes to run a marathon, and of just how much faster the elite runners are from the bulk of the participants. Finally, the tool must allow users to compare finish times between different demographic groups, both in terms of averages and extremes.

Design

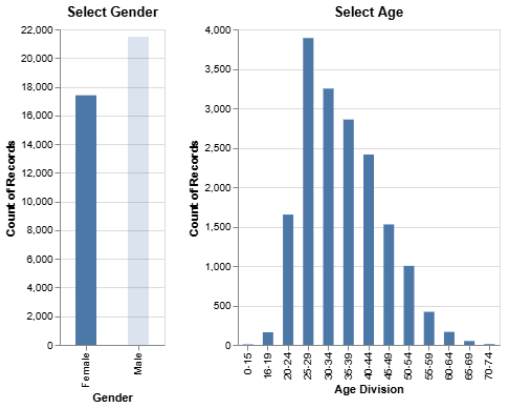
The design phase of this project included several prototypes and revisions. Early on, I decided not to include the Country variable in the analysis and focus on the relationships between the three variables of Gender, Age, and Finish Time. My first idea was to use pie charts to display the distributions of gender and age, but some age groups were very small compared to others, so bar charts proved a better option. For the finish times, I thought to have one large histogram with all of the times in the background, overlaid by a selected demographic sample. This turned out to be ineffective, as the samples were often too small to show features when compared against the entire dataset. The histogram therefore displays only the finish times from the selected sample. This, however, makes it difficult to directly compare finish times between groups. In order to effectively convey the relationship between age/gender and finish time, I decided to add a scatterplot that showed the average time as well as the fastest time from each demographic group. Below is a screenshot of the final design, implemented using Altair, as well as a link to the interactive tool:

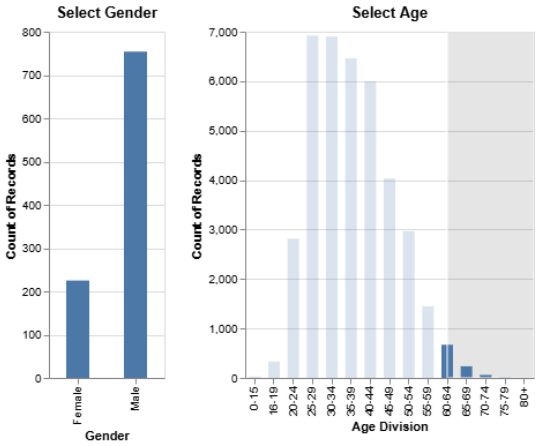


Link to html file: https://github.com/edejongh1/testrepo/blob/main/Marathon.html

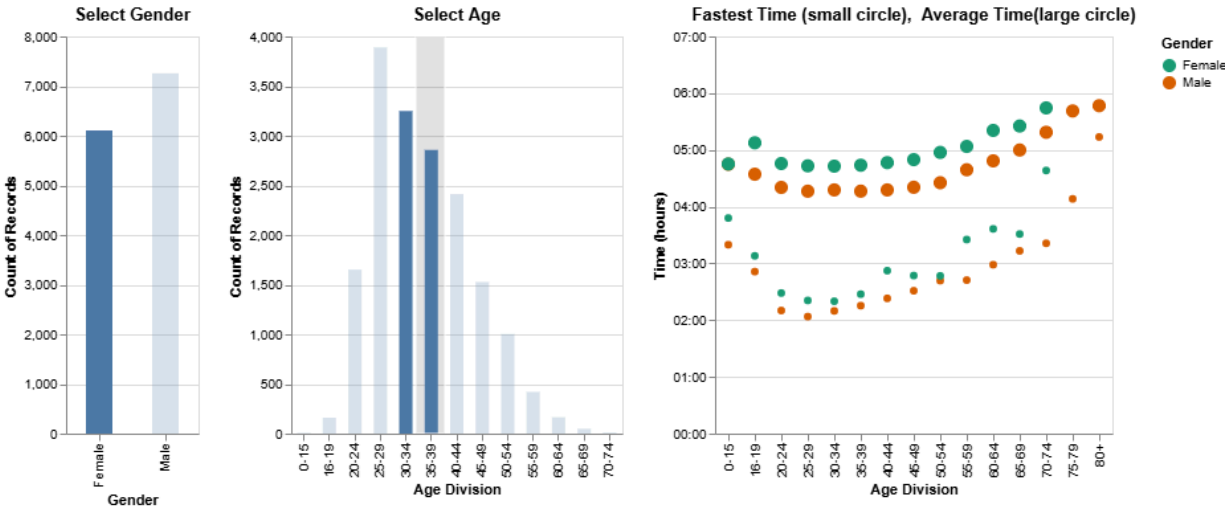
The first two bar charts show the distributions of gender and age among the population of runners. By default, the whole dataset is selected and the distribution of finish times is displayed in the histogram on the bottom. The scatterplot on the top right shows the relationship between age and finish time, color-coded for gender, with large circles indicating the average time for each group and small circles indicating the fastest time in the group.

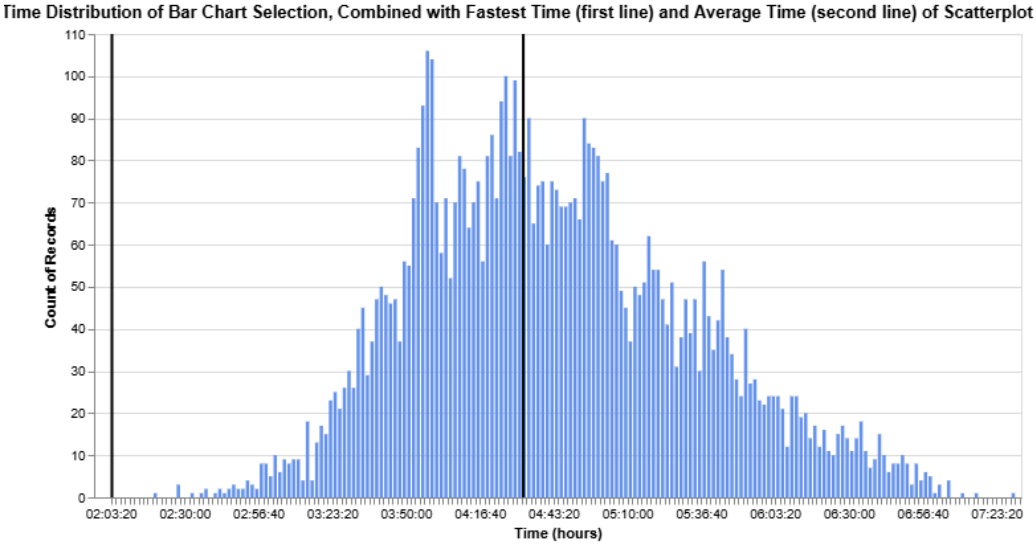
Gender can be selected by clicking on one of the bars in the first chart, and a range of ages can be selected by clicking and dragging in the second chart. The two charts update each other when selections are made, which allows for some exploration of the demographic distribution. One can see, for example, that female runners skew younger than male runners, and that runners aged 60+ are largely male:



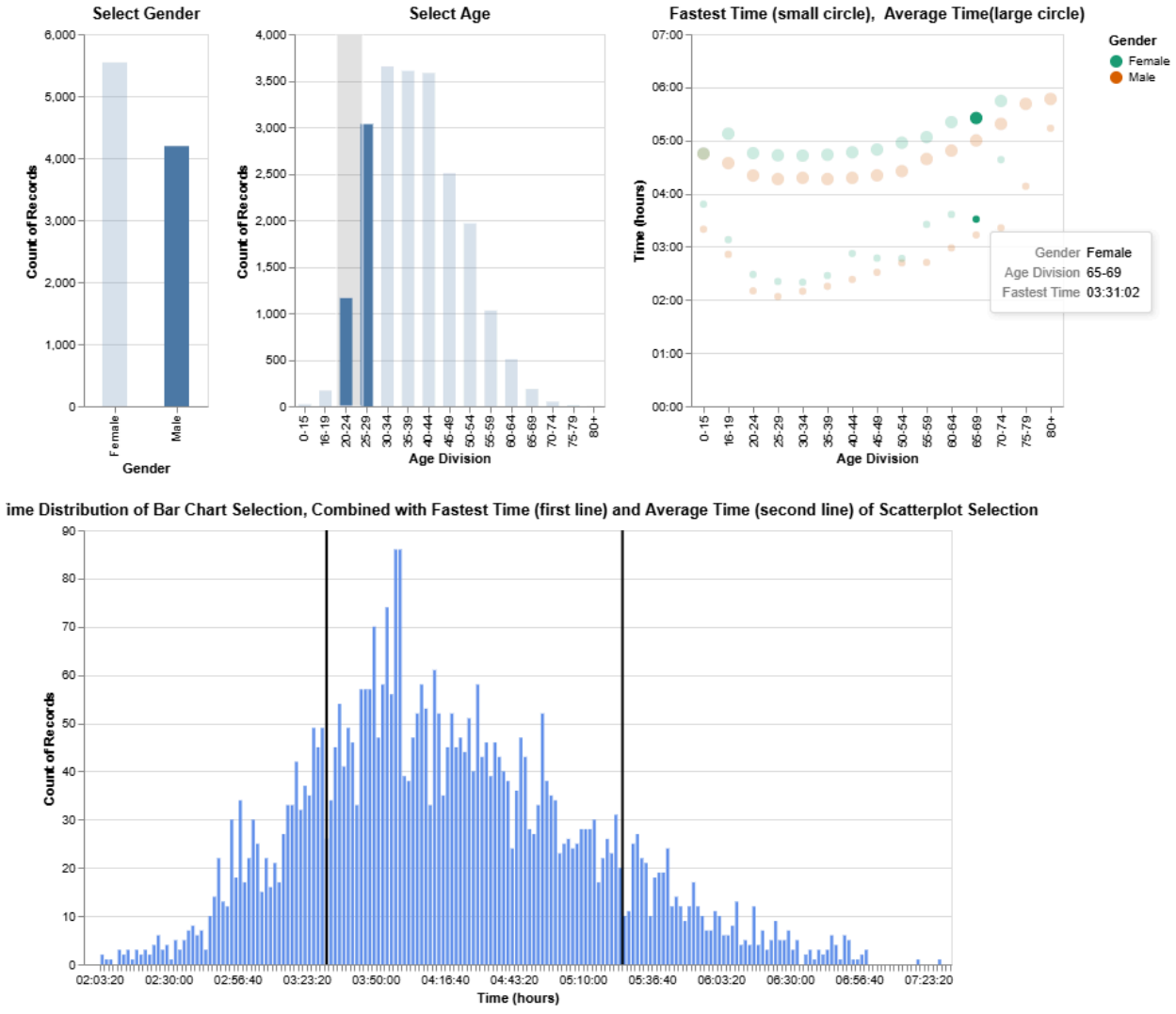


Once a selection is made, the histogram updates. For example, the histogram below shows the time distribution for female runners aged 30-39:

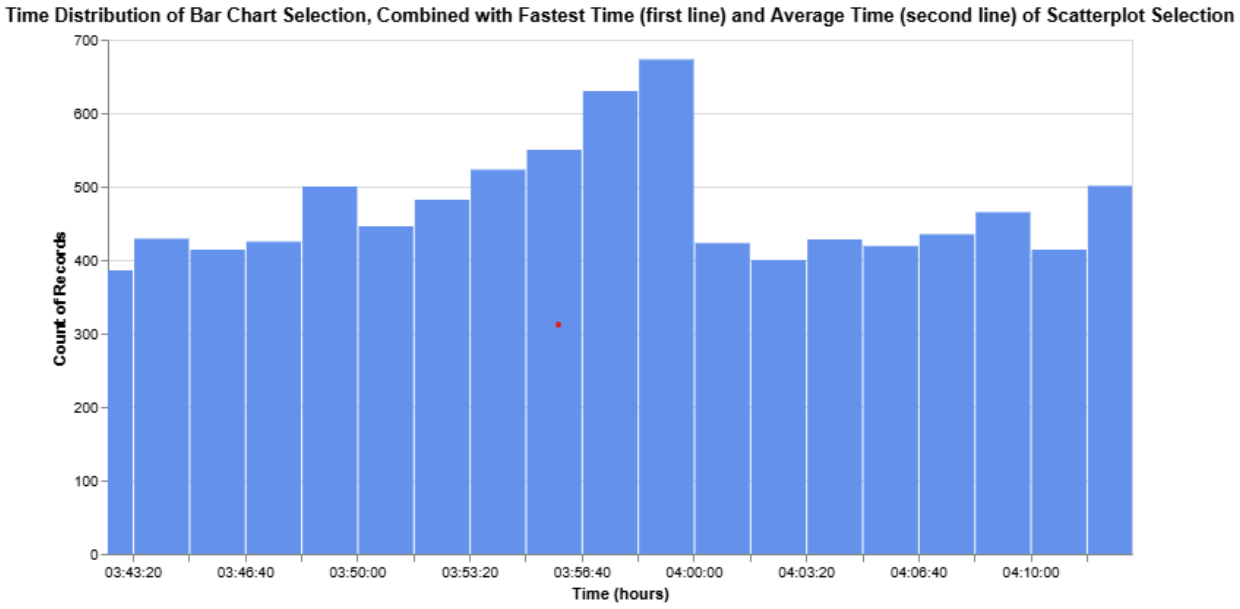




The scatterplot comes with a tooltip feature that shows the user the gender, age group, and exact time by hovering the cursor over a point. In addition, selecting a point on the scatterplot causes the fastest time and average time from the selected group to be displayed as vertical lines on the histogram, to be compared with the separately selected time distribution. This allows for some revealing comparisons. For example, the fastest woman in the 65-69 age group was quite a bit faster than most men aged 20-29 (though on average, the men were faster):



Finally, the histogram comes with a zoom-in feature for examining the time bins in more detail. One can see, for instance, that a large drop-off occurs right before the 4-hour mark when looking at the whole dataset:



Evaluation

To evaluate this visualization tool, I recruited three people to try it out and conducted qualitative interviews about their experience. My questions focused on how easy they found the tool to use, what insights they gained from using it, and what improvements they would like to see. The three participants included an avid marathon runner, a scientist with experience using data visualization, and a family member with comparatively little experience in this domain.

The responses I received were generally positive, although one participant thought the functionality of the charts should be communicated more clearly with more descriptive labels and instructions. Once they became familiar with the functionality, the participants found the tool intuitive and insightful. One person noted the drop-off structure before the 4-hour mark of the histogram and postulated that many runners were aiming for that milestone. Participants also expressed surprise at how slowly performance dropped off with age, and at how wide the distribution of finish times was overall.

All three participants had suggestions for new features to add to the tool, which I took as a positive sign that the tool kept them engaged and wanting to learn more. One suggestion was for a feature that would allow users to select a range of finish times from the histogram and display the gender/age distributions for that range, in a reverse of the current functionality. Other suggestions included a way to directly compare histograms of two different groups, an option to show the median time on the scatterplot instead of the mean, and an option to display the time in terms of minutes per mile

instead of total finish time. For functional improvements to the existing tool, participants suggested more well-defined tick marks on the histogram and labels for the vertical lines indicating mean and minimum values.

Conclusion

This data visualization project successfully created a tool for insightful analysis of marathon racing results. It effectively presents an overview of the data, showing a diverse population of runners with a wide range of results, while also allowing users to drill down and examine subsets of the data in more detail. One of the key insights communicated through this visualization is that while - on average - a clear relationship exists between gender/age and finish time, each demographic group contains runners of a wide range of ability levels, such that one cannot easily predict an individual's race time based solely on their age and gender. Future iterations of this tool would aim to add options to present the data in more ways that might lead to better understanding for some users, such as presenting race times in minutes per mile (or km) or miles per hour. Feedback from user interviews has provided suggestions for new comparison features that could also be included. Future evaluations would be needed to determine whether these new features enhance the user experience or add unnecessary complexity to the tool.