Process Book

Ultra Running Insights

Ethan Pedersen u1054170 ethan.pedersen@utah.edu
Jackson Dean u1100004 jackson.dean@utah.edu

Git Project Repo: ultrarunning-insights

Overview & Motivation

Running performance data is of personal interest to our group because we both have a background in the sister sport of track and field. The ultra-distance running community has seen high growth over the past 20 years relative to shorter road races, making it more relevant than ever.

From a data perspective, we believe ultra running(50km and greater) is an interesting and unique avenue to explore for several reasons. For one, the length of the races provides for more complex data fields to examine than shorter distances such as the marathon or sprint distances, such as rest times, food intake, and the effects of ascents/descents in elevation. Ultra races themselves are unique and have high variation in distance, so comparison between races will be complex. The sport also has an interesting gender aspect, as it is already known that the gender gap in performance shrinks as the distance increases, and on average women perform better than men in distances greater than 195mi (Source).

Related Work

The website RunRepeat.com has published several articles (like this one) analyzing ultrarunning data to find various interesting trends. Additionally, there are various race data collection sites across the internet, but none of them do much with their impressive databases other than allow runners to look up their results. For example, see runnercard.com, itra.run, and ultra-marathon.org.

The “wave” bar chart visualization was thought of independently but then inspired after-the-fact by the Table Lens visualization shown in lecture.

Questions

One main relationship we set out to explore is between gender and running performance, and how the gender gap is affected by various factors. We wanted to present these relationships through fun and interesting visuals. Other groups have published findings of this nature before, but with very research paper-esque “boring” graphs.

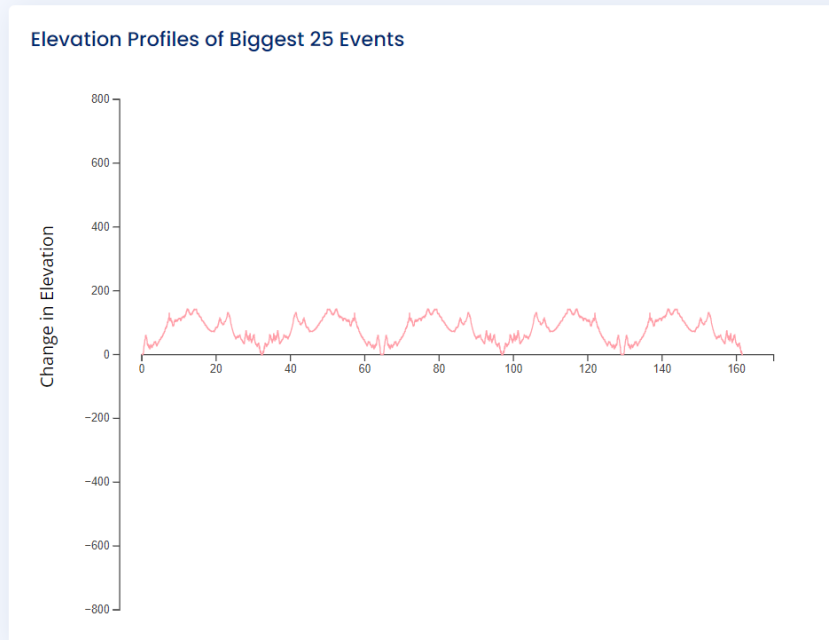
Data

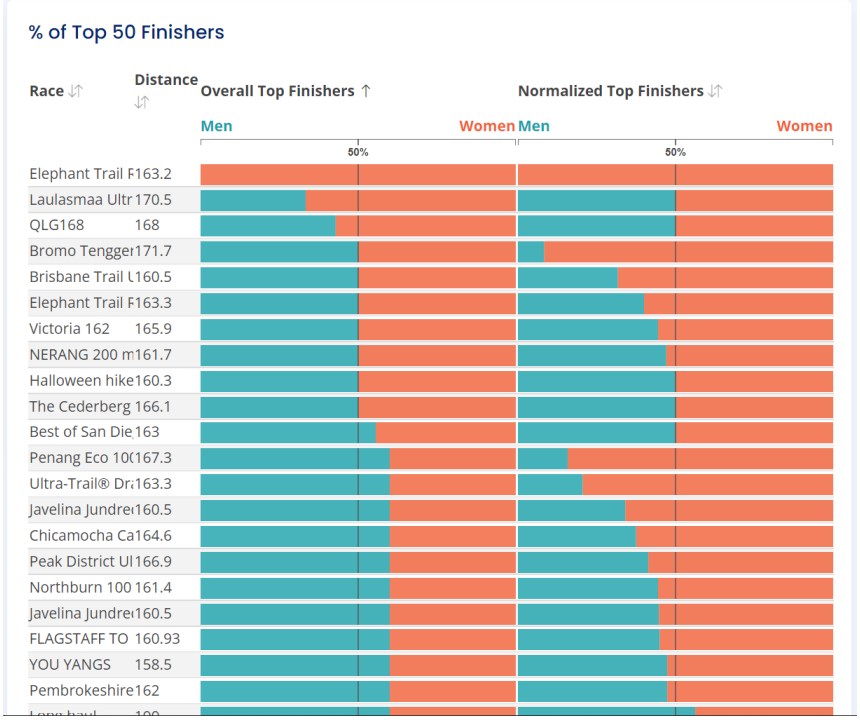
Exploratory Data Analysis

Some basic data analysis was done using spreadsheet software to examine the raw dataset. From other analyses that we were aware of, we knew what trends to expect in general. This exploration was primarily to find which of these trends we would be able to extract from the particular dataset that we had access to.

We also used exploratory analysis to determine which races to get course profiles for. Because it involved manual searching online for individual course profiles, we decided to limit it to 25 - 50 races. We have been able to find .gpx files for almost every race we've searched for, and we can use the .gpx data to generate a course elevation change profile.

Design Evolution





Implementation

Evaluation

==========

Original Proposal document below

Background and Motivation

Running performance data is of personal interest to our group because we both have a background in the sister sport of track and field. The ultra-distance running community has seen high growth over the past 20 years relative to shorter road races, making it more relevant than ever.

From a data perspective, we believe ultra running(50km and greater) is an interesting and unique avenue to explore for several reasons. For one, the length of the races provides for

more complex data fields to examine than shorter distances such as the marathon or sprint distances, such as rest times, food intake, and the effects of ascents/descents in elevation. Ultra races themselves are unique and have high variation in distance, so comparison between races will be complex. The sport also has an interesting gender aspect, as it is already known that the gender gap in performance shrinks as the distance increases, and on average women perform better than men in distances greater than 195mi (Source).

Project Objectives

One main relationship we would like to explore is between gender and distance running performance, and how the gender gap is affected by various factors. We would like to present these relationships through fun and interesting visuals. Other groups have published findings of this nature before, but with very research paper-esque “boring” graphs. We also expect that as we become familiar with the data, we may find other interesting relationships. For example, I expect that we will find correlations between performance and location, elevation, runner nationality, season, weather, and any of the other dozens of factors affecting a given race.

We would also like to be able to compare the relative difficulty of races compared to one another. ‘Difficulty’ could be determined by elevation ascent / descent, average temperature, dropout / completion rate, average of top course performances in terms of speed, and more.

Data

So far we have found one comprehensive data set of races from 2012 to 2021 pulled from the International Trail Running Association (ITRA) that will provide what we need, freely available here on Github. This dataset is recent and large enough to get us through our main goals for this project. ITRA is an international organization that looks to grow the ultra running community and increase diversity in the sport. ITRA maintains up-to-date results for nearly all known ultra distance races, so it may be worth it for us to learn how to scrape both more recent and older data from their collection, as 2020 and 2021 data may be unusual due to the pandemic. ITRA also maintains a Performance Index for the community, which compares athletes against each other by weighting finish times with race course characteristics.

If possible, it would be interesting to find health metric data for individual athletes such as mile or km splits, heart rate, and calorie intake. This may be infeasible for several reasons as such data is published at the discretion of individuals, but it is possible that this has been collected for top finishers at the most popular races in recent years.

Data Processing

The primary dataset we currently have is already nicely processed in CSV format. A small handful of entries (< 75) are missing race distance or other vital data fields, for which we will have to fill in manually with some research.

Since our dataset is static, we will do most of our processing a single time and append the results to the dataset itself rather than programmatically calculating at runtime. We will likely perform some additional data processing to extract derived data points (i.e. distance between

aid stations) and to retrieve corresponding data from other sources (i.e. starting elevation retrieved from geographical data sources based on the provided city and country and historical weather data from the race days). We will also collect course maps, a large number of which are available online in GPX format.

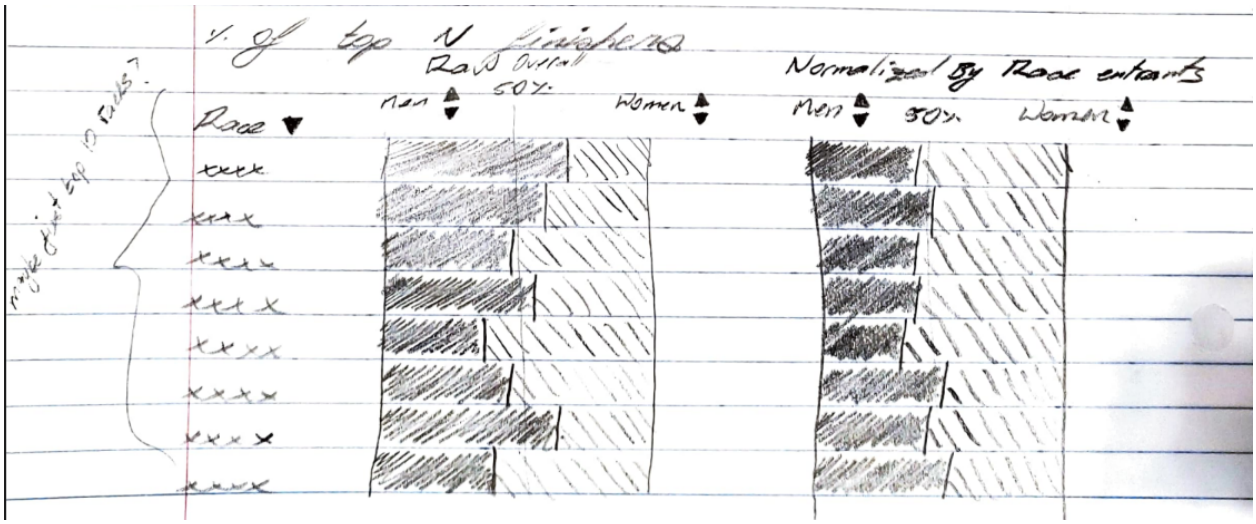
Visualization Design

Sketches

We knew from the start that there would be no single visualization that would convey all of the information we wanted, so we opted for a dashboard of some kind that would collect multiple cimpler charts. Most of our prototyping process was deciding on a variety of visualizations that would communicate a variety of interesting data.

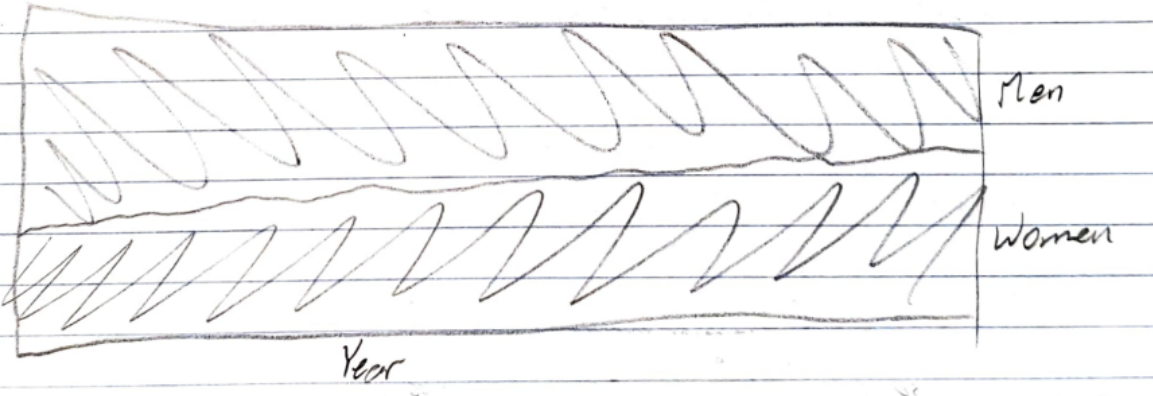
Following is a variety of chart ideas that we came up with.

These three graphs are used to show the difference between men and women performance which was the original aspect that we wanted to explore with this dataset.



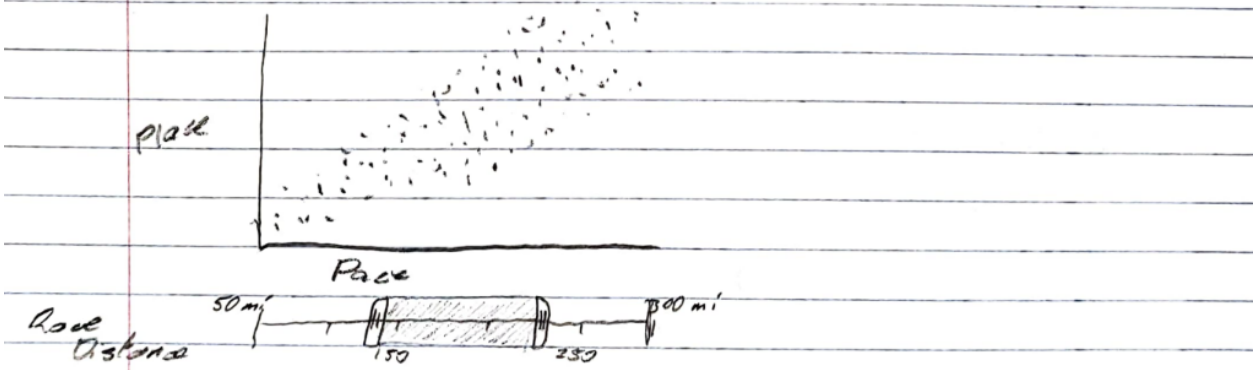
Percentage of women finishing in top N places by year

(overall row & normalized by entrants)

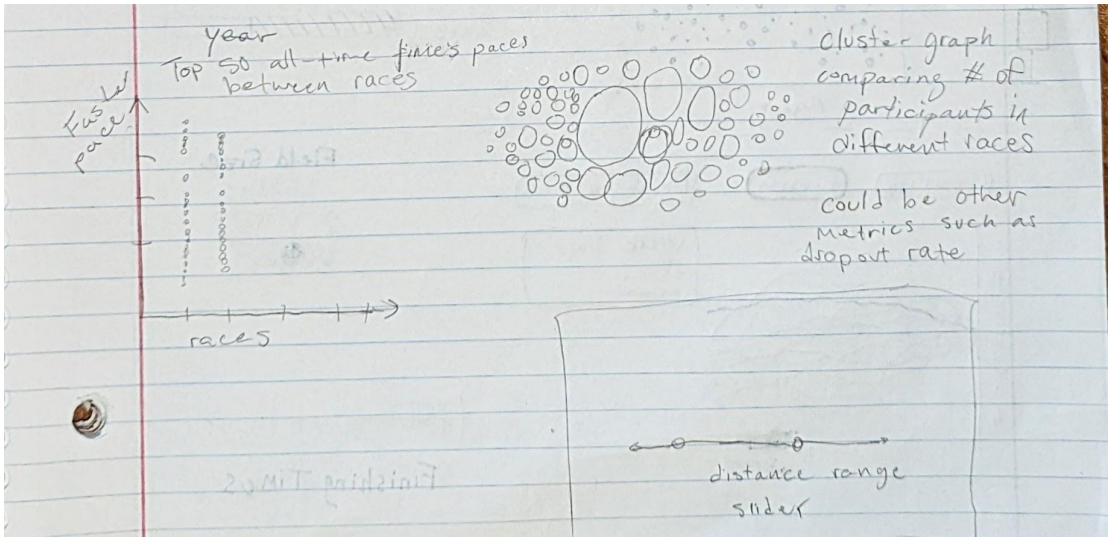


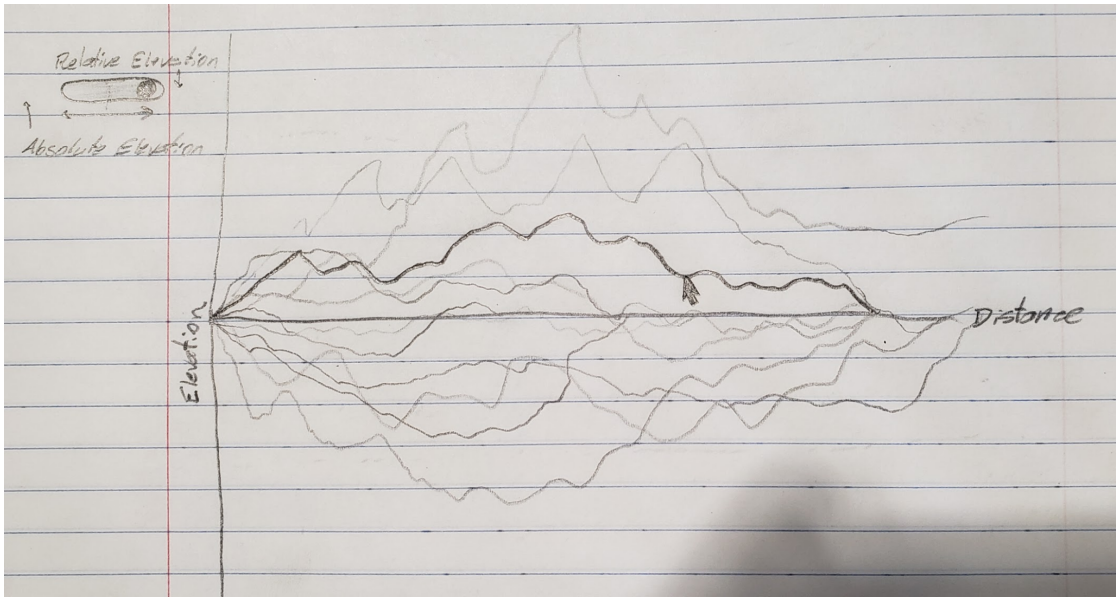
For a given selected race, simple scatterplot of finish times. Shaded by gender? Or 2D by gender

Sister bar for distance with graph showing pace vs place. Points shaded for men vs women

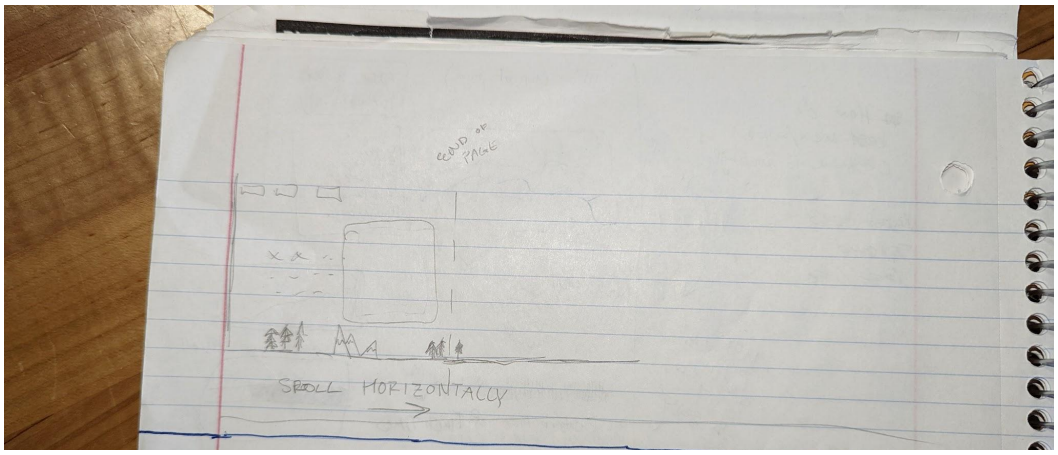


These charts show general race statistics, not necessarily related to gender.



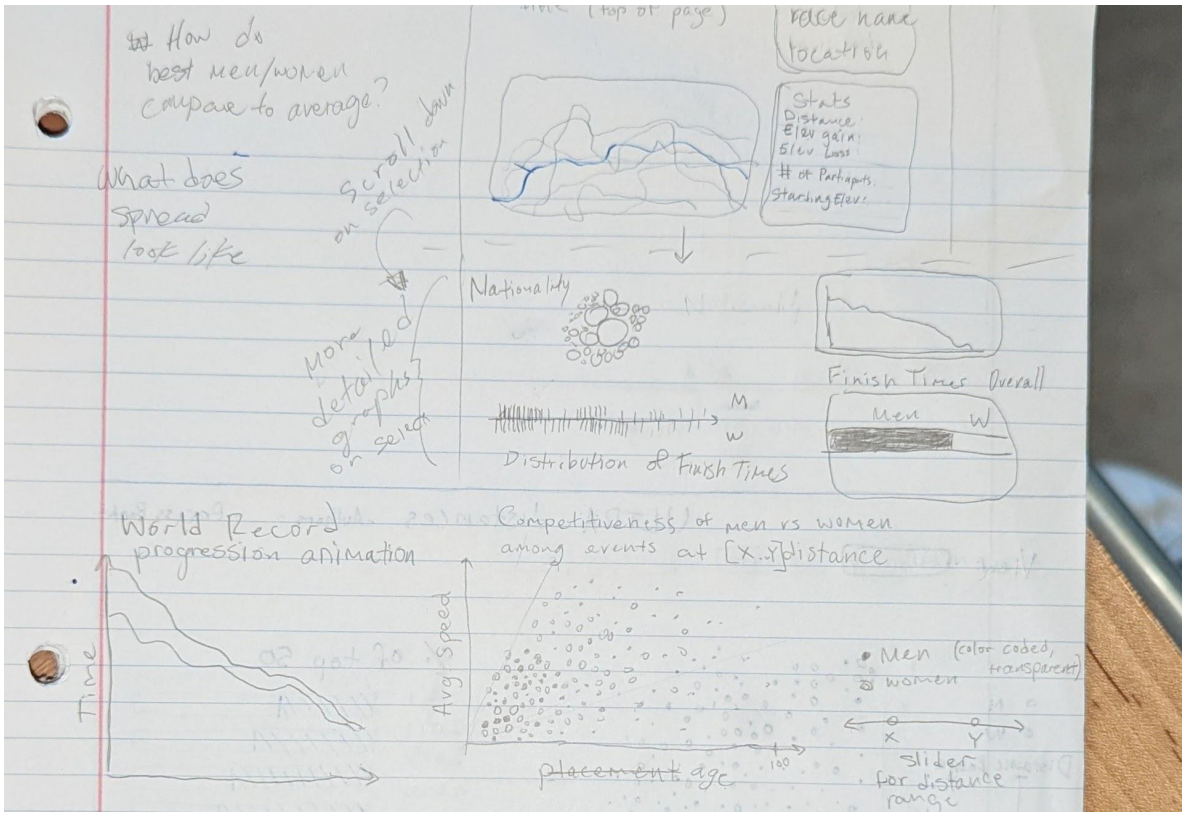


A basic idea for a side-scrolling layout.

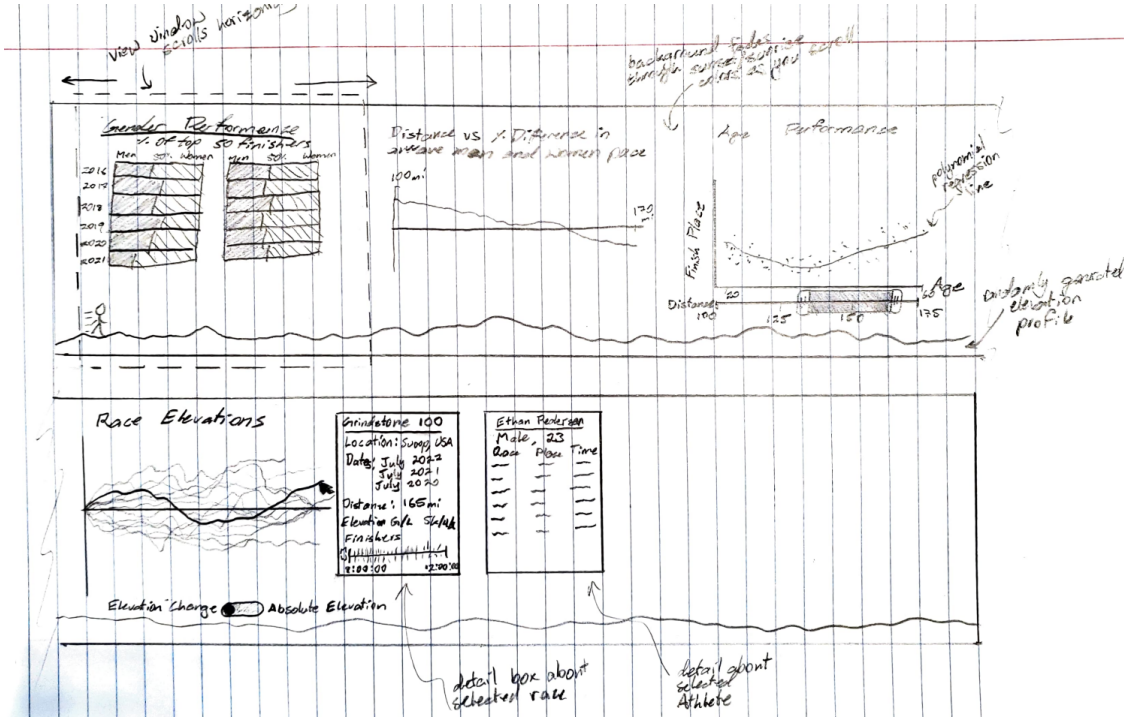


Prototype Layouts

A data dashboard style layout



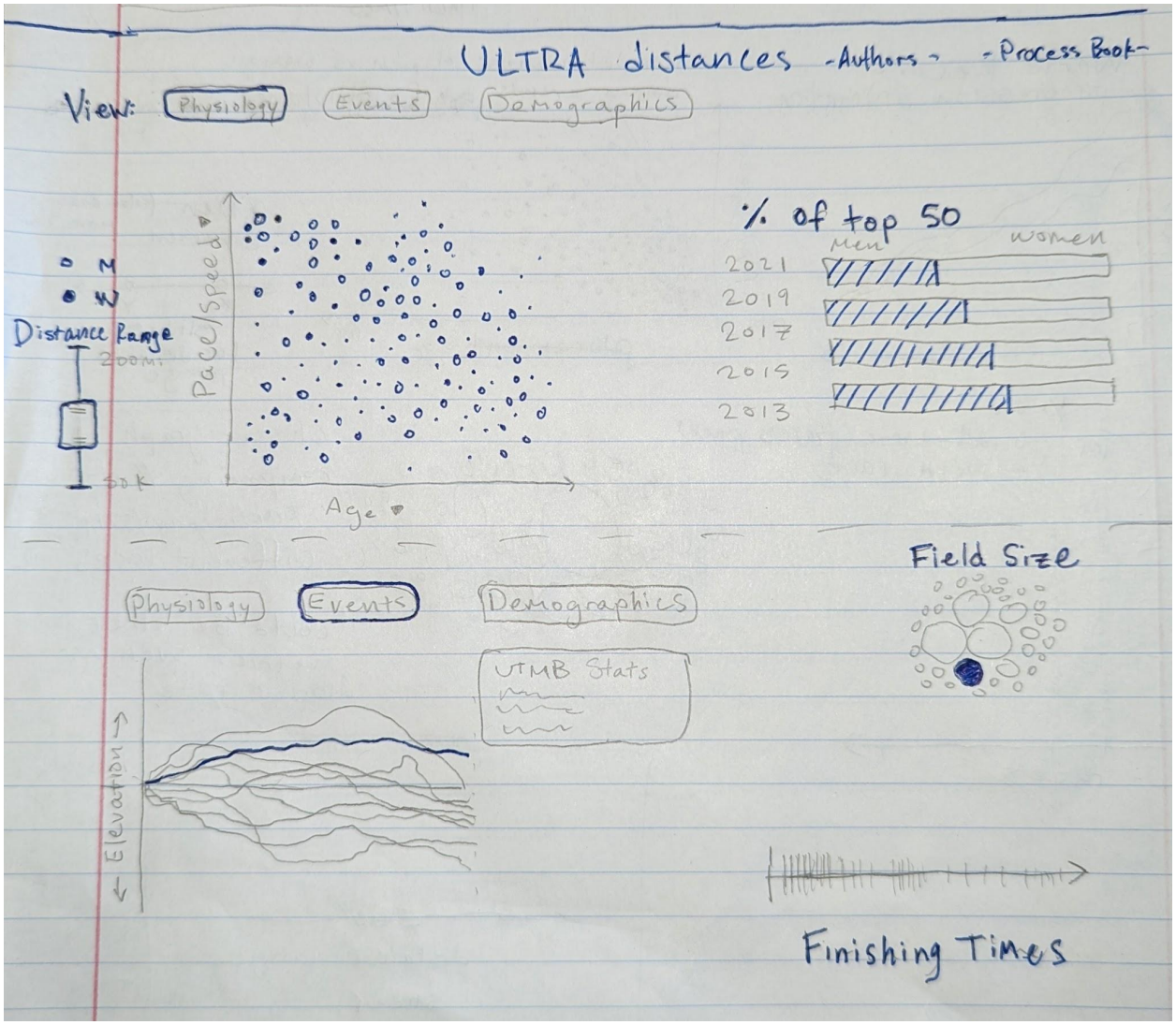
A side-scrolling site that presents separate visualizations one at a time.



Final Design Idea

Although we really liked a horizontal idea with ‘running’ animations, we ultimately went back to the dashboard style single page. What separates this as our final design is a tabular-like interface to bring different data categories together. What helped us achieve this was broadening our scope so that the categories make sense as separate tabs. In this design, you can see the top view of graphs associated with athlete physiology, while the bottom part separated by the bottom line shows what it will look like with the Events view.

The Demographics view will look similar, but we are currently unsure of which visualizations will be included on that page since we have not finished acquiring country demographics data.



Features

Must-Haves

- Percent bar graph of men vs women top finishers per race, with both absolute percentages and percentages normalized by participants.
- Interactive line chart with elevation profiles of races overlaid on each other. This will have filter selections on which races are displayed, and allow you to toggle whether it shows relative elevation change profiles or absolute elevation. Biggest / most popular races (UTMB, Western States 100, etc.) will be highlighted initially. Selection of one will populate other panels with statistics of that race.
- Bubble graph that uses area to compare races on different metrics to highlight the highest in each metric such as participation field size and completion rate.

Optional

- Chart comparing nationality of participants vs country's average elevation, GDP, and other economic metrics to visualize how location affects opportunity of participation and performance.
- Animation showing change in world records at different ultra distances such as 50k, 100k, 100mi, etc over time between genders.
- Visualize the average of top 50 or so course times between races at the same distance

Project Schedule

1. Oct 17th
    - Project proposal (All)
2. Oct 24th
    - Get datasets loaded in and cleaned/processed; (Jackson)
    - Find geocoded data for courses; (Ethan)
    - General website layout (All)
3. Oct 31st: Create the must-have graphs and have display features ready
    - Men vs Women bar graph (Jackson)
    - Bubble chart with variable race metrics (Jackson)
    - Elevation profile overlay (Ethan)
    - Individual race statistics panel (Ethan)
4. Nov 7th
    - milestone document (All)
    - add interactions for main graphs (All)
5. Nov 11th: Milestone due
6. Nov 14th
    - Add animations (All)

    - Polish main charts and interactions (All)
7. Nov 21st
    - 100% completion of must-haves (All)
    - Country demographics charts (Jackson)
    - World records animation (Ethan)
8. Nov 28th
    - Wrap up, polish site and prepare for presentation (All)