EPFL COM-480 Data Visualization Project Process Book

Joey Zenhäusern, Sung Lin Chan, Xiangzhe Meng

Overview & Motivation

The Lausanne Marathon is an annual marathon race held in Lausanne since 1993. This road running takes place in autumn (October) and the 20 km of Lausanne takes place in spring (April). Parallel to the marathon, a half-marathon, a 10 km, a 10 km walking and Nordic walking, and a mini-marathon race of 4 km are organised. The Lausanne Marathon is one of the largest annual sporting events of the Canton de Vaud, and attracts up to 2,500 tourists each year.

The race starts in Lausanne, Place de Milan, and then follows the road along the Lake Léman until La Tour-de-Peilz, before returning to Lausanne via the same route. The arrival is located in front of the Olympic Museum fountain. The half-marathon (21 km) starts from La Tour-de-Peilz (since 2000) and ends in Lausanne.

At the very beginning, the initial reason for why we chose this topic is that we are 3 marathon enthusiasts. We are curious about this well-known marathon competition and the average performance of the runners who participates in it.

Hence, our project aims at visualizing the results of the Lausanne Marathon of the past 16 years ranging from 2002 to 2017. In particular, we would like to make hidden patterns apparent, such as distributional differences regarding age groups, genders or experienced compared to inexperienced runners. The data also provides some interesting possibilities for map creation, since the hometown of runners, and in some cases, the nationality of the runners, are available to us.

Our visualization could be useful for past and prospective runners to get a overview of where they stand with their performance, for DataSport who collects time measurements at swiss sport events to help enhance their data presentation, and even for event sponsors who would like to know more about the runners who participate in this event.

Dataset

Data loading and pre-processing

Our data scraping notebook can be found by following this link:

https://github.com/zenhaeus/cs-480/blob/master/notebooks/data-scraping.ipynb

As the data is not available to download in a single file, we had to scrape the website of "Datasport", the company that provides time measurement for Swiss broad sport. The data in original form can be found by clicking on the respective years here: https://en.lausanne-marathon.com/palmares/2017/

Parsing and preparing the data proved to be a little challenging, since over the years the column names, the number of columns and the way the data was displayed, changed significantly. In a first step we had to split the columns by identifying white-space gaps over entire pages of data. This was necessary because the columns were not separated by single characters, but instead by variable amounts of spaces, as can be seen here: https://services.datasport.com/2016/lauf/lamara/alfaa.htm

In a next step we normalized the column names, as some columns were German or otherwise inconsistent. We also had to split up columns and move some data between them to get a clean column with the cities where the participants are from. Next we only kept columns that were shared across all years and merged all of the years into one dataframe. The column 'catégorie', which specifies which race a participant took part in, also had to be normalized. We decided to only keep male and female participants of the following races: 10km, half marathon, marathon. We also split this column up into two columns: gender and race distance.

Finally we used the google geocode api to look up the coordinates of all the participants' home cities in the dataset, in order to visualize where the participants are from on a map. After some additional cleaning, such as dropping rows that could not be geocoded or that contained otherwise invalid data and normalizing birth years, the data should be ready for visualization.

Data description

Our dataset contains the results of the Lausanne Marathon races over the past 16 years. Although there are lots of different kinds of races, we took only the marathon (42km), half-marathon (21km) and the 10km race into consideration.

The raw data contains:

- **Name**: Runner name
- **Birth year**: Runner birth year
- **Age**: Runner age
- **Age group**: Runners are classified into 4 age groups: 20-30, 30-40, 40-50, Other
- **Gender**: Runer gender (male & female)
- **Race year**: Race year when the runner participates in the competition
- **Category**: There are 3 kinds of races: 42 stands for marathon, 21 stands for half-marathon, 10 stands for 10km short-distance marathon
- **Rank**: Runner rank. As the participants are classified into different groups, the final rank is computed in terms of the runner's gender, age group and race category instead of only considering race category.
- **Time**: Runner's run time during the competition in the form hh(int) : mm(int) : ss(float)

- **City**: Hometown city of the runner
- **Canton**: Hometown canton/department of the runner:
- **Country**: Runner's nationality
- **Latitude**: Latitude of the runner's hometown
- **Longitude**: Longitude of the runner's hometown

Here is an example form of our data after cleaning and preprocessing;

A	B	C	D	E	F	G	H	I	J	K	L	M	N
name	birth year	age	age group	gender	race year	category	rank	time	city	canton	country	latitude	longitude
Aarskog Kay-Morten	1981	36	30-40	male	2017	10	676	15:48.7	Oteren	Troms	NO	69.2555679	19.8844188
Haaland Arne Kristoffer	1981	36	30-40	male	2017	21	179	37:17.0	Oteren	Troms	NO	69.2555679	19.8844188
Annen-Lamard Sandra	1977	40	40-50	female	2017	21	1	22:42.5	Ecublens	VD	CH	46.5296363	6.5615252
Bach Philippe	1988	29	20-30	male	2017	10	108	40:17.6	Ecublens	VD	CH	46.5296363	6.5615252
Balsiger Patrick	1956	61	other	male	2017	10	25	48:42.1	Ecublens	VD	CH	46.5296363	6.5615252
Barrera Roca Adria	1992	25	20-30	male	2017	42	136	43:41.3	Ecublens	VD	CH	46.5296363	6.5615252
Benedetti Marta	1989	28	20-30	female	2017	10	33	44:10.4	Ecublens	VD	CH	46.5296363	6.5615252
Berret Emmanuelle	1985	32	30-40	female	2017	21	84	50:41.7	Ecublens	VD	CH	46.5296363	6.5615252
Bilardo Giuseppe	1960	57	other	male	2017	10	125	48:25.3	Ecublens	VD	CH	46.5296363	6.5615252
Boffin Niall	1997	20	20-30	male	2017	21	109	36:27.0	Ecublens	VD	CH	46.5296363	6.5615252

Exploratory data analysis

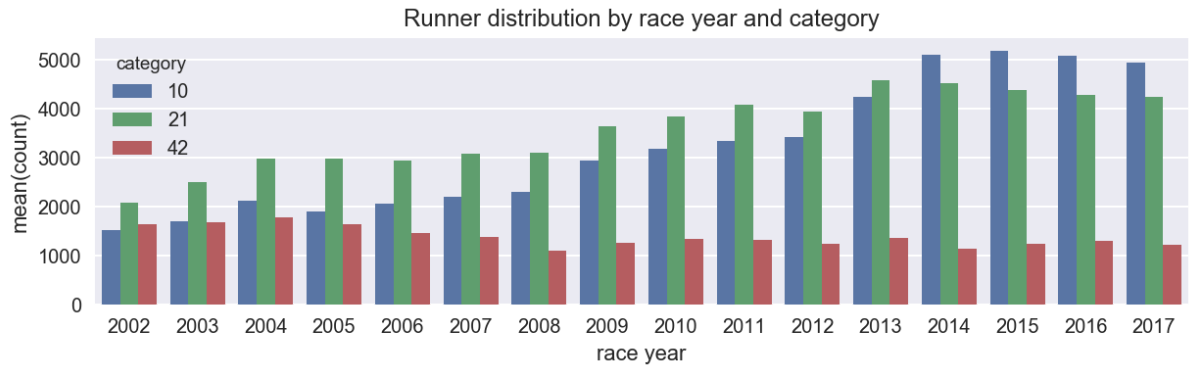
Our data analysis ipython notebook can be found by following this link:

https://github.com/zenhaeus/cs-480/blob/master/notebooks/Data%20analysis%20for%20visualization%20designs.ipynb

After cleaning and pre-processing our data, we did some exploratory data analysis in order to pinpoint our design direction and construct some more rational and interesting visualization to tell our data story.

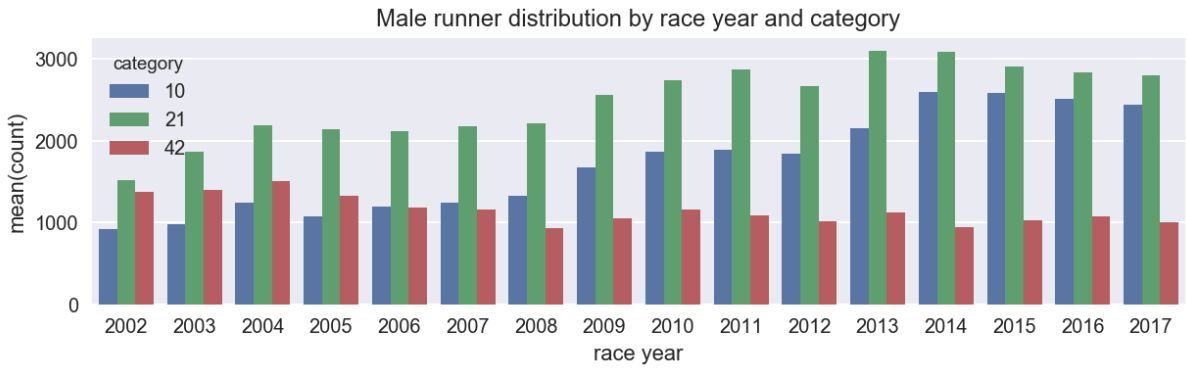
Runner distribution by race year and category

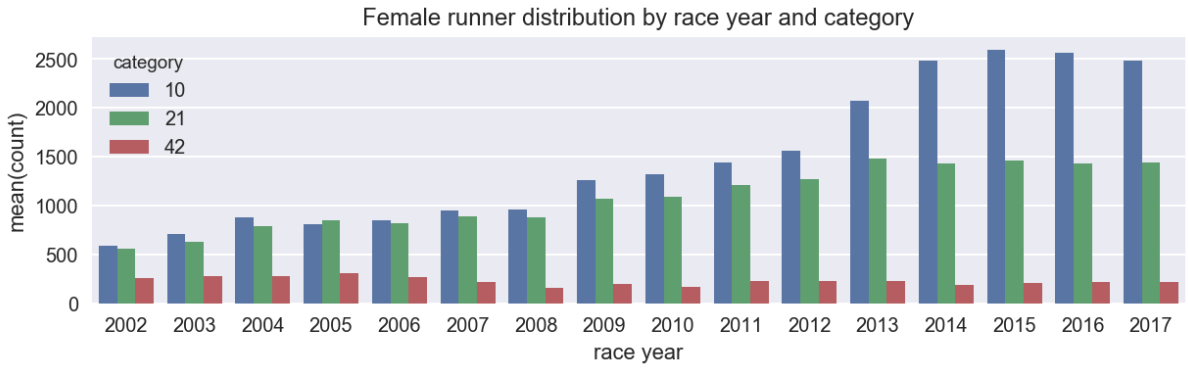
First of all, we have a look at the runner distribution of the 3 different races for each year. From the bar plot below, we can easily found that the number of runners who participate in 10km and 21km marathon increase a lot from 2002 to 2017. While less and less people participate in the 42km marathon over year, although the variation is not evident.



Then, we plot separately the female and male runner distribution of the 3 races in order to find out if the variation tendency and the runner distribution of the 3 races are the same as the plot above for both males and females.

From the plots below, we can roughly conclude that the variation tendency of all the races are very similar. However, the runner distribution of the 3 races are totally different.

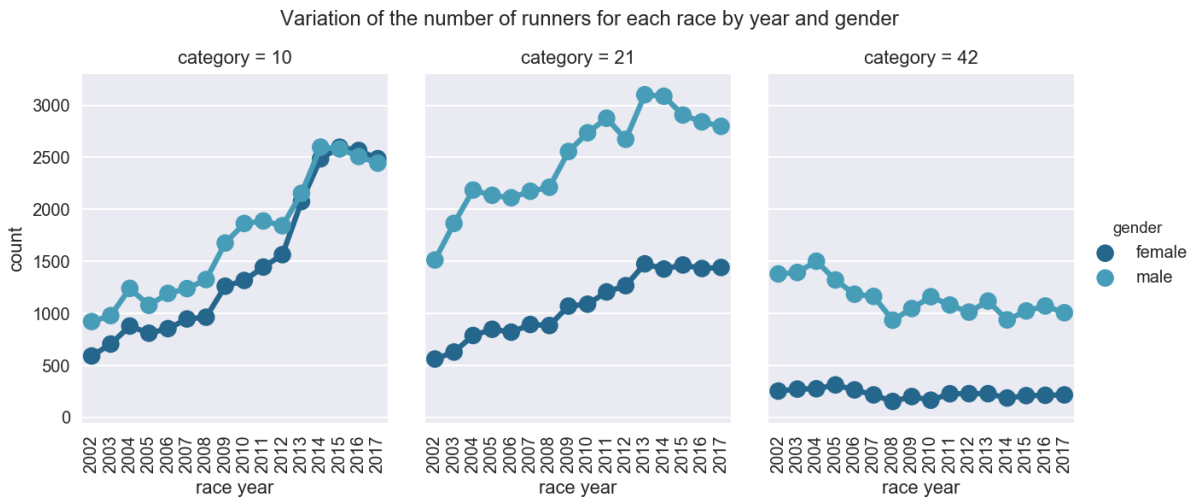


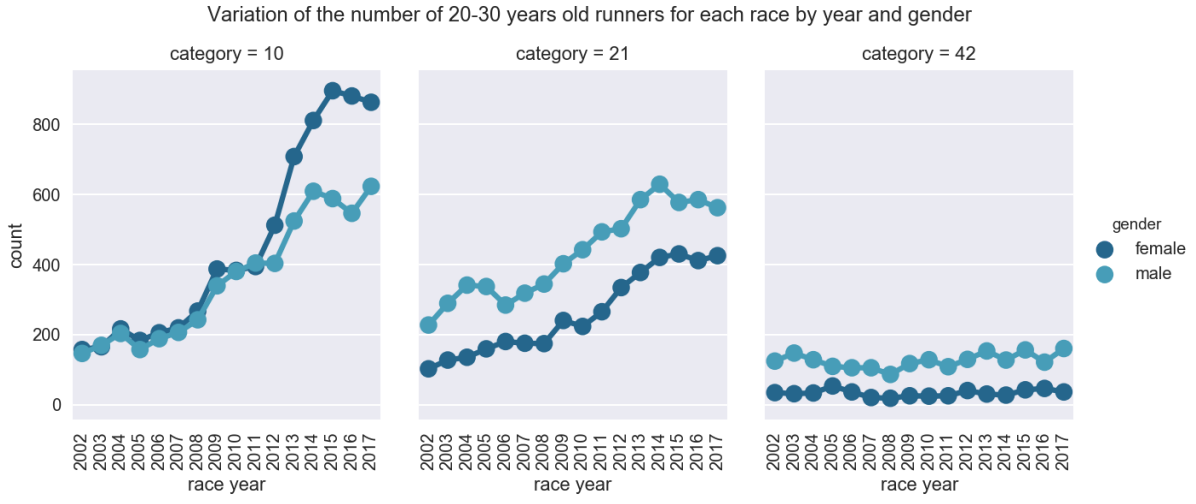


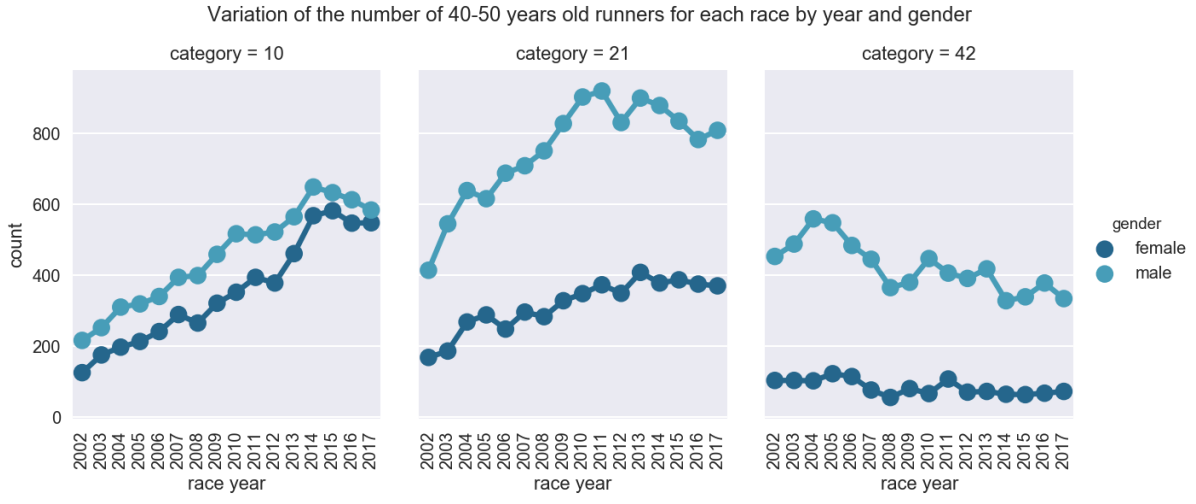
Variation of the number of runners for each race by year and gender

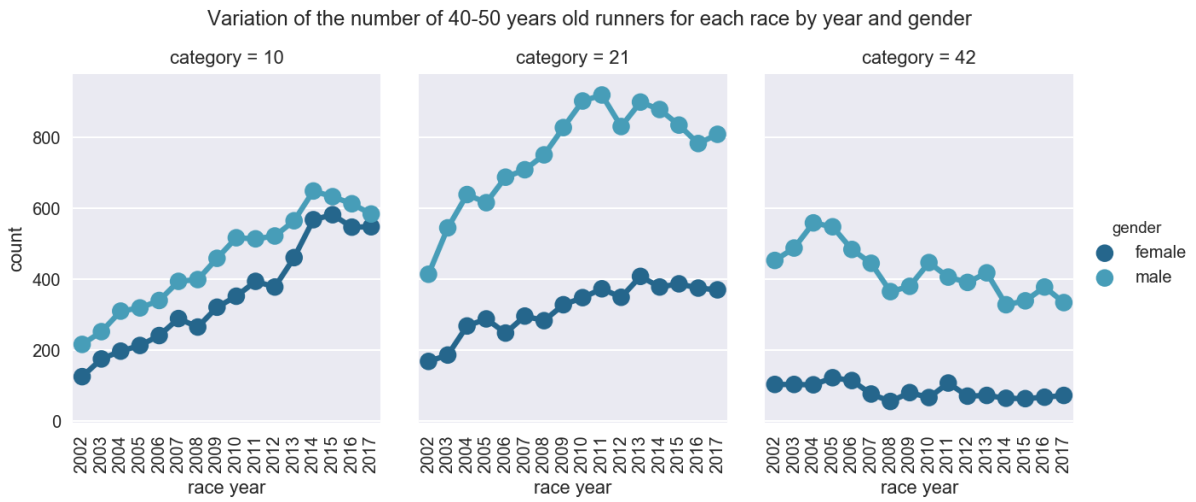
In this part, we mainly focus on the variation of the number of male and female runners for each race.

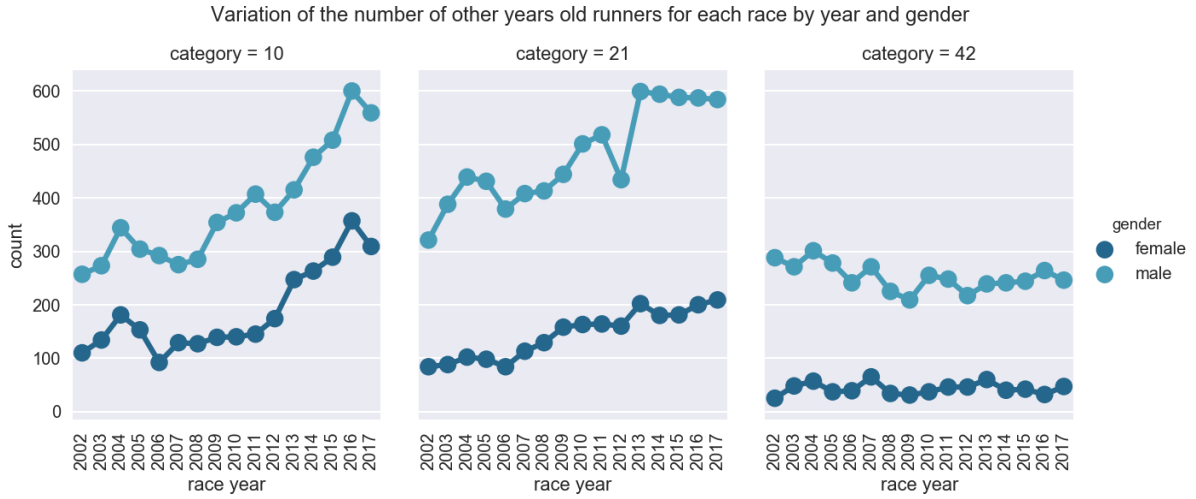
Then, we divide all runners into 4 different age groups: age between 20 and 30, age between 30 and 40, age between 40 and 50, Child & Senior(Other) so as to observe the variation of the number of male and female runners in different age groups for each race.


Variation of the number of runners for each race by year and gender
category = 10
category = 21
category = 42
count
race year
gender
female
male


Variation of the number of 20-30 years old runners for each race by year and gender
category = 10
category = 21
category = 42
count
race year
gender
female
male

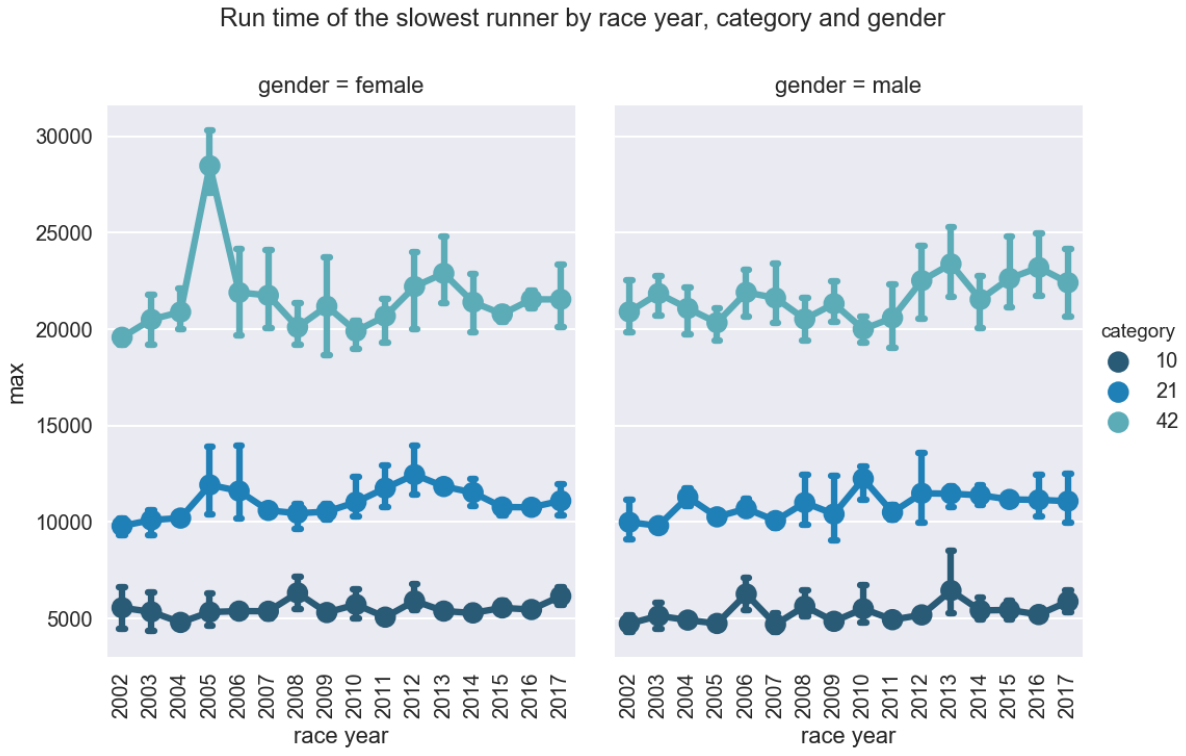

Variation of the number of 40-50 years old runners for each race by year and gender
category = 10
category = 21
category = 42
count
race year
gender
female
male

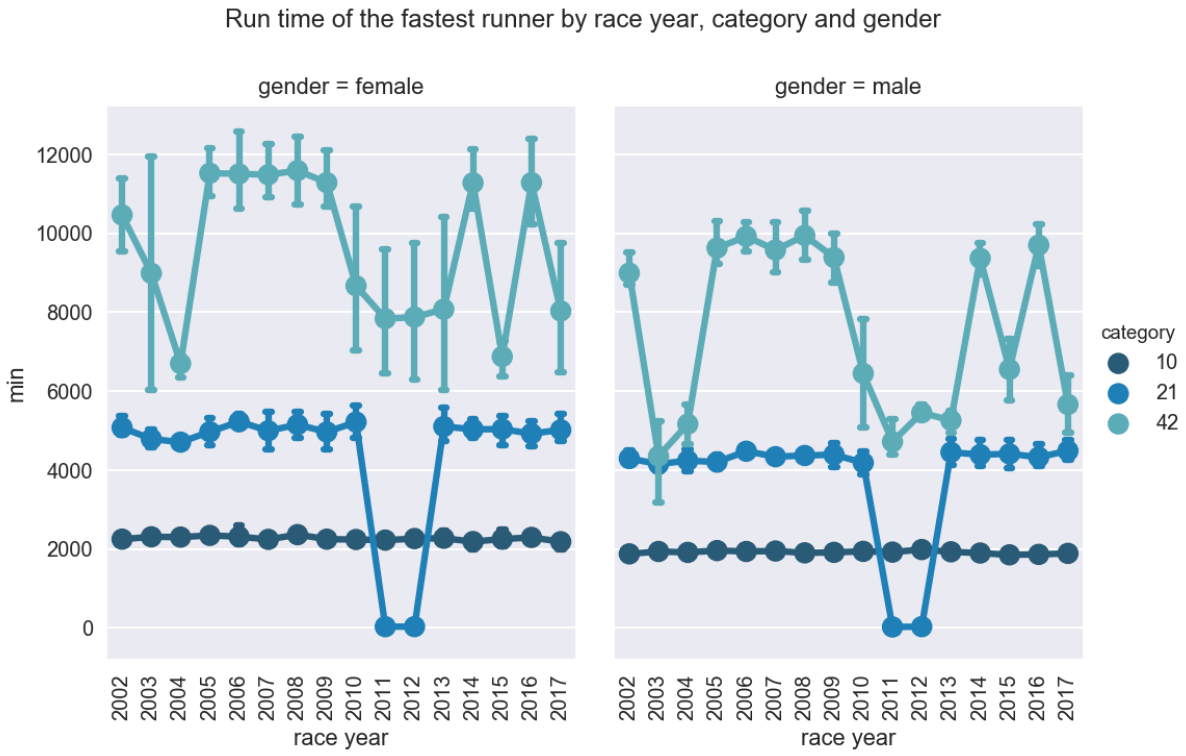


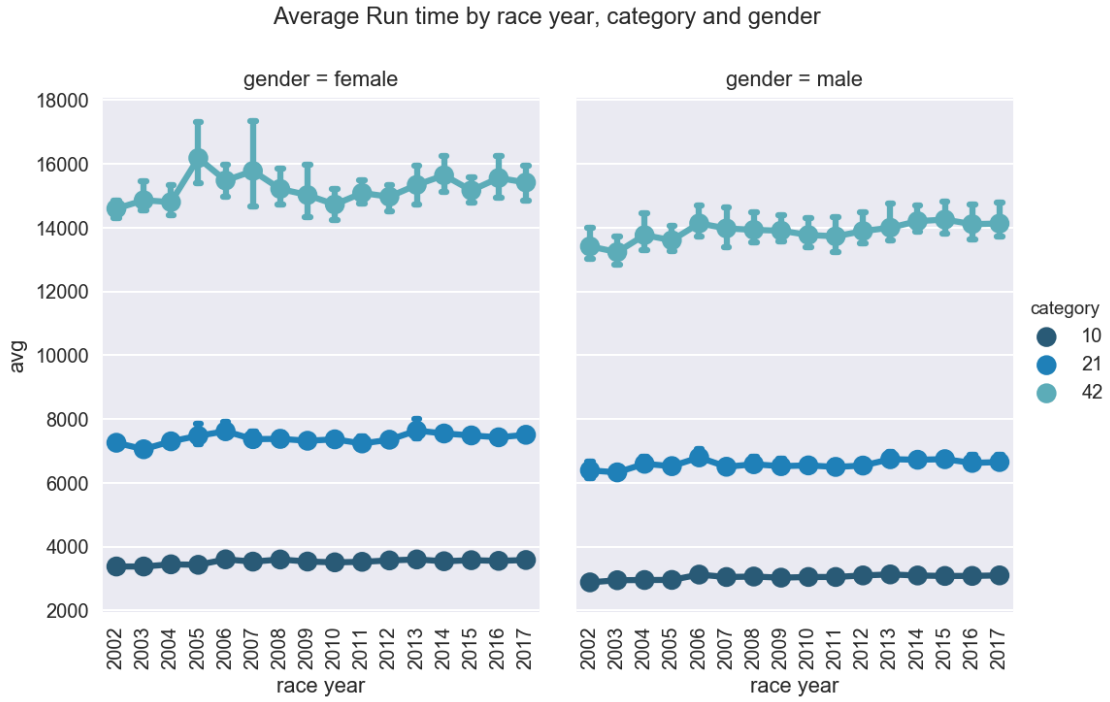


Performance comparison over year

In this part, we plot the variation of the fastest/slowest runner performance and the average run time over year by race and gender. We can easily observe that the variation of the best and worst performance change a lot over year. The average run time, however, doesn't vary so much.


Run time of the slowest runner by race year, category and gender
gender = female
gender = male
max
30000
25000
20000
15000
10000
5000
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
race year
category
10
21
42


Run time of the fastest runner by race year, category and gender
gender = female
gender = male
min
12000
10000
8000
6000
4000
2000
0
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
race year
category
10
21
42



Hometown city of top runners distribution

In this part, we compute the number of cities that the top 10 female and male runner of each age group come from scaling from 2002 to 2017. From the result, we can find that those best runners doesn't always come from the same city. Hence, we are curious to know the best runner's geographical distribution over year and expect to observe more insights with the help of map.

```
From 2002 to 2017,
        For age group 20-30,
                Top 10 male runners of 42km marathon come from 109 different citys.
                Top 10 female runners of 42km marathon come from 105 different citys.
                Top 10 male runners of 21km marathon come from 103 different citys.
                Top 10 female runners of 21km marathon come from 97 different citys.
                Top 10 male runners of 10km marathon come from 83 different citys.
                Top 10 female runners of 10km marathon come from 89 different citys.
        For age group 30-40,
                Top 10 male runners of 42km marathon come from 124 different citys.
                Top 10 female runners of 42km marathon come from 108 different citys.
                Top 10 male runners of 21km marathon come from 113 different citys.
                Top 10 female runners of 21km marathon come from 106 different citys.
                Top 10 male runners of 10km marathon come from 89 different citys.
                Top 10 female runners of 10km marathon come from 95 different citys.
        For age group 40-50,
                Top 10 male runners of 42km marathon come from 139 different citys.
                Top 10 female runners of 42km marathon come from 128 different citys.
                Top 10 male runners of 21km marathon come from 125 different citys.
                Top 10 female runners of 21km marathon come from 114 different citys.
                Top 10 male runners of 10km marathon come from 90 different citys.
                Top 10 female runners of 10km marathon come from 87 different citys.
        For age group other,
                Top 10 male runners of 42km marathon come from 134 different citys.
                Top 10 female runners of 42km marathon come from 120 different citys.
                Top 10 male runners of 21km marathon come from 115 different citys.
                Top 10 female runners of 21km marathon come from 95 different citys.
                Top 10 male runners of 10km marathon come from 90 different citys.
                Top 10 female runners of 10km marathon come from 87 different citys.
```

Country distribution

At the very beginning, we find that there are 93 countries which have participated in the Lausanne marathon competition for at least one time since 2002, but most of the countries don't have so many participants. With curiosity, we decide to have a look at the country distribution of this competition in order to obtain some more profound insights.

In this part, we mainly do a statistical analysis aim at the countries which have at least 100 runners participated in the competition over the past 16 years.

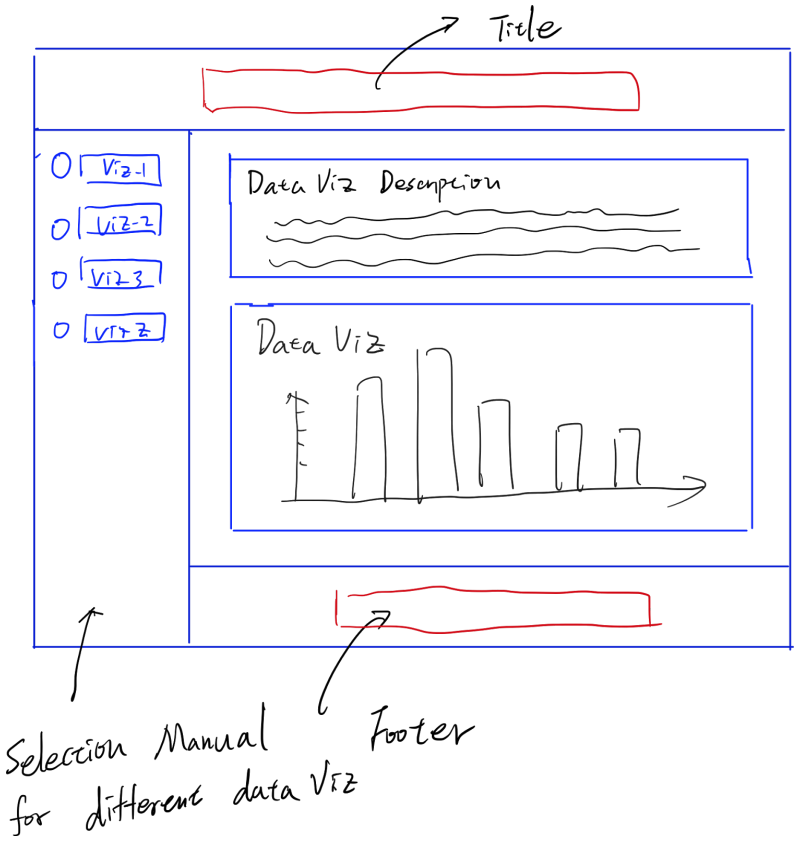
country	count runner	count year	count male	count female	best rank
CH	101041	16	65230	35811	1
FR	17640	16	12761	4879	1
US	4035	16	2699	1336	1
DE	1580	16	1079	501	1
IT	969	16	705	264	1
GB	819	16	505	314	1
BE	607	16	415	192	1
CA	514	16	310	204	1
ZA	420	16	310	110	6
JP	357	16	191	166	1
NL	260	16	180	80	4
DK	201	16	105	96	1
ES	162	16	125	37	2
IN	143	16	104	39	1
SE	136	14	86	50	1
CZ	130	14	95	35	3

Visualization Design

Website

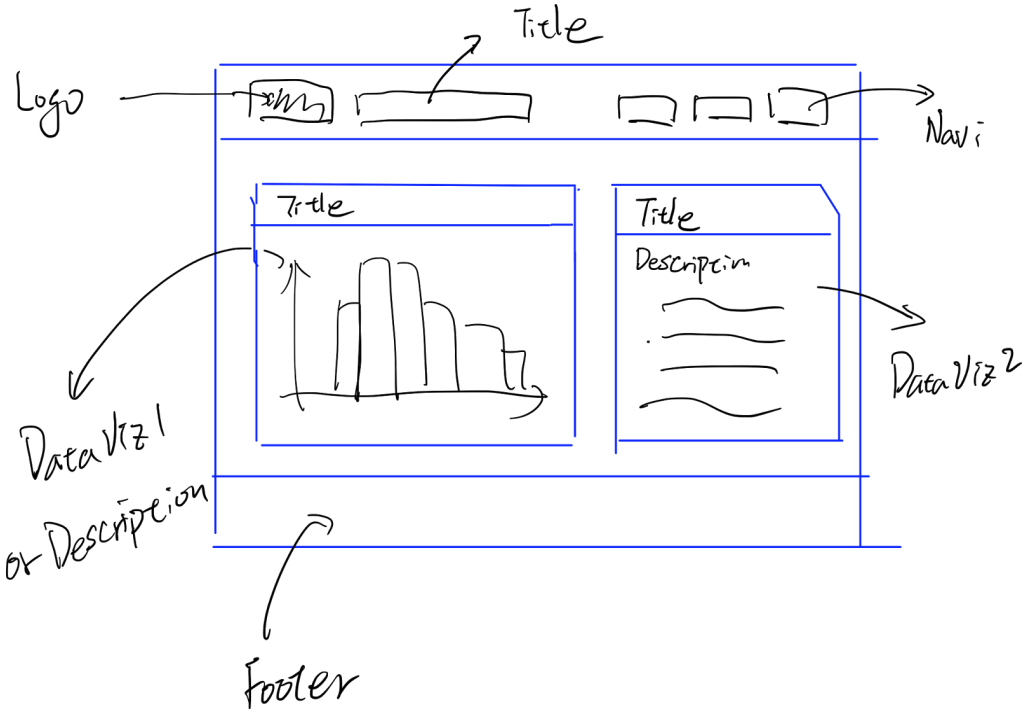
11/11

We think our goal is to present the visualizations. Therefore, we use the very simple layout for our main page. It only contains a title on the top, a navigation bar on the left and the footer at the bottom. The rest of the space is remained for the main character , data visualization.



12/ 10

We find that the previous version of the navigation bar occupied too much space on the left. Therefore, we decided to move the navigation bar to the top. This move could give more space for the main visualization.

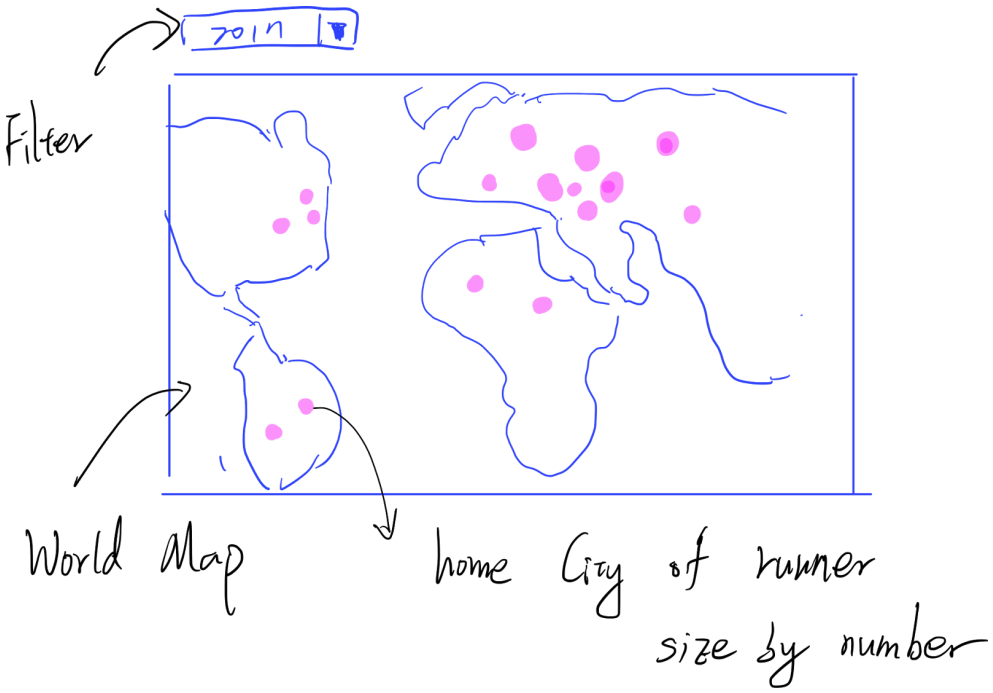


Data Visualization

1. Locality

11/ 13

Our original idea of the visualization is try to show the distribution of the participants in the marathon by years. The map is definitely our choice. And we would like to show the different sizes of circles proportional to the number of participants from each city in order to give the insight of the locality of runners.

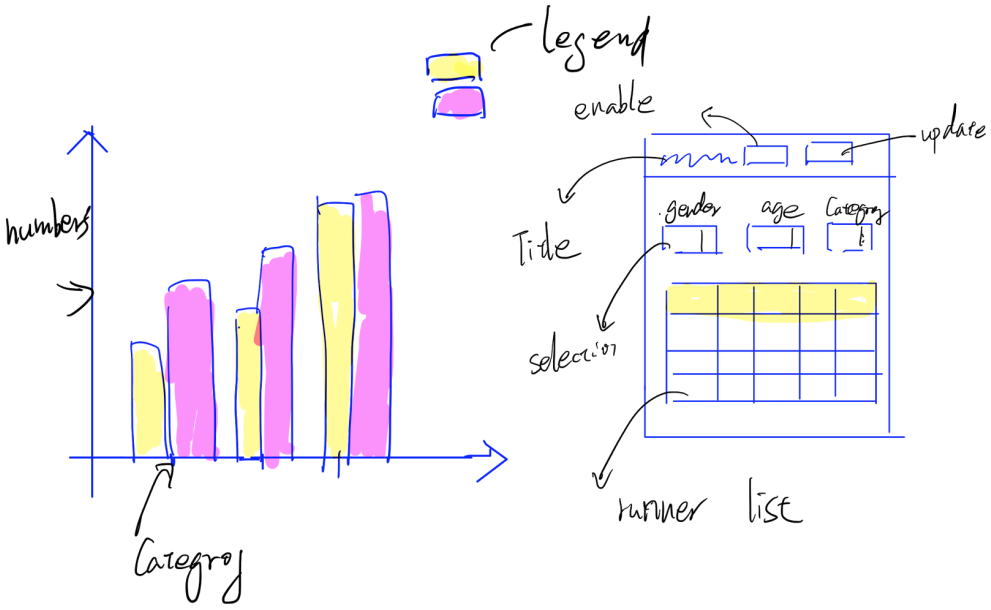


11/20

We implemented a prototype and observed an unexpected result. The big circles are intense around Switzerland. And when zooming out the map is not easy to give the user the information of the runner. We discussed with TA about this visualization. He suggested that we can try heap map instead of showing the circles on the map. However, after implementing the heap map, the problem wasn't solved. The area of switzerland is still very "hot" and can't be informative.

12/02

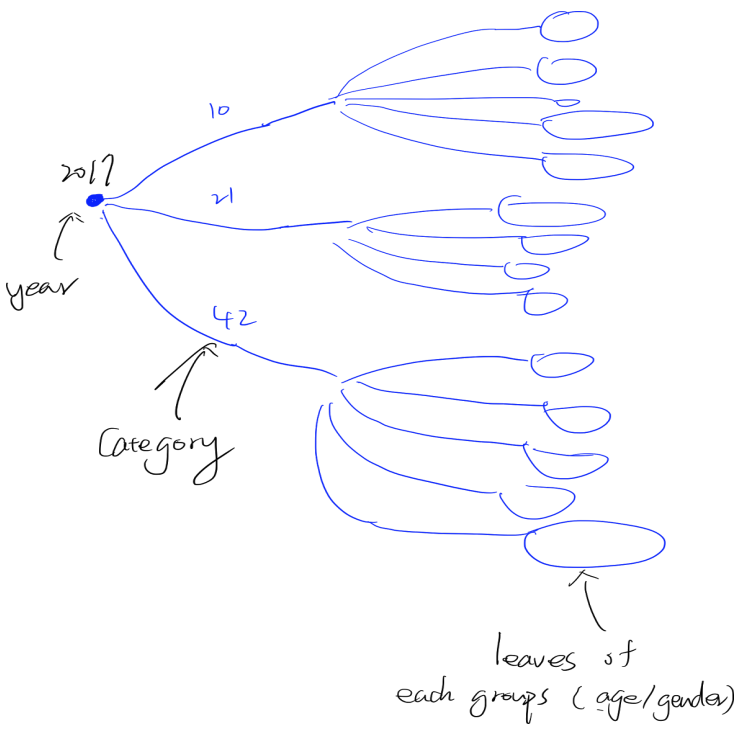
In order to give more information to the user, we decided to link our map to two different graphs. One is the group bar chart with number of participants which outer x-axis is the category and the inner x-axis the gender. Moreover, we want to add hover effect on each bar, and display the distribution of each group. The other is the list of the top 10 runner of each race, and show their home city with the different notation on the map. This feature could show the insight of where the best players came from.



2. Race Description

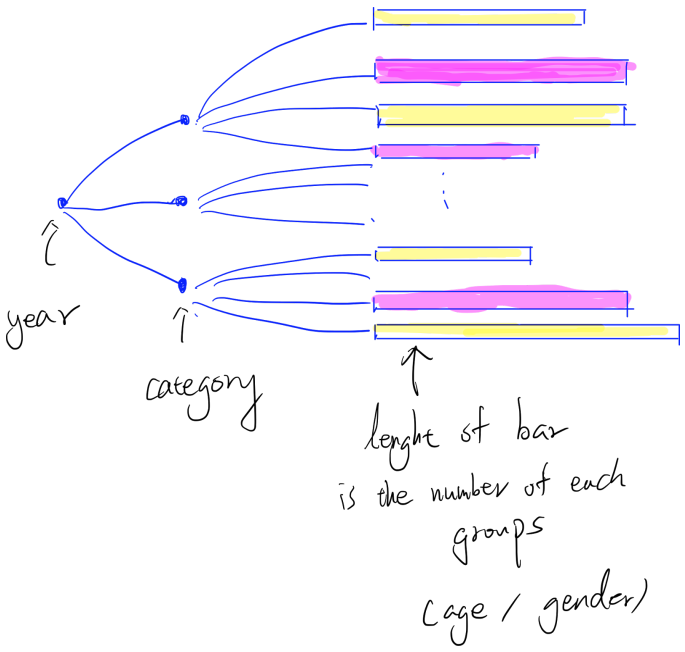
11 / 13

Initially, the purpose of this dendrogram is to give user a brief perception about the structure of Lausanne marathon competition since the structure and the way they computed the final rank is not intuitionistic and more or less complicated. From our point of view, using dendrogram is the most efficient way to present the competition structure. And we would like to present the end of the dendrogram like the leaves of the tree, an oval like shapes. The size of the leaves reflect the number of the people in the group.



11/20

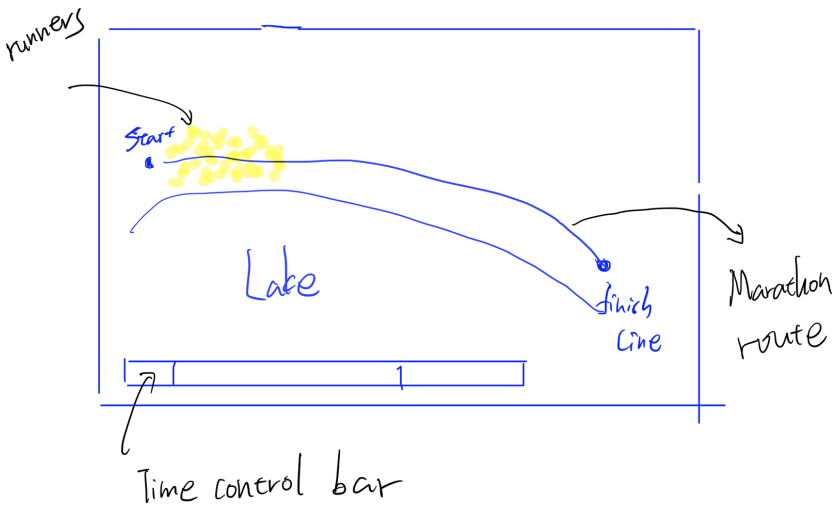
We decided to change the leavest to the rectangular bars. Because, it is more obvious to see the difference among the group. The size of leaves is much harder to distinguish the differences and truly reflect the the figures of the data.



3. Speed

11/ 13

For this visualization, we try to show all the runner's positions on the map, and see how the group of people moves along the time. We plan to add select menu, let the user chose the different group, ane observe the movement of crowd.

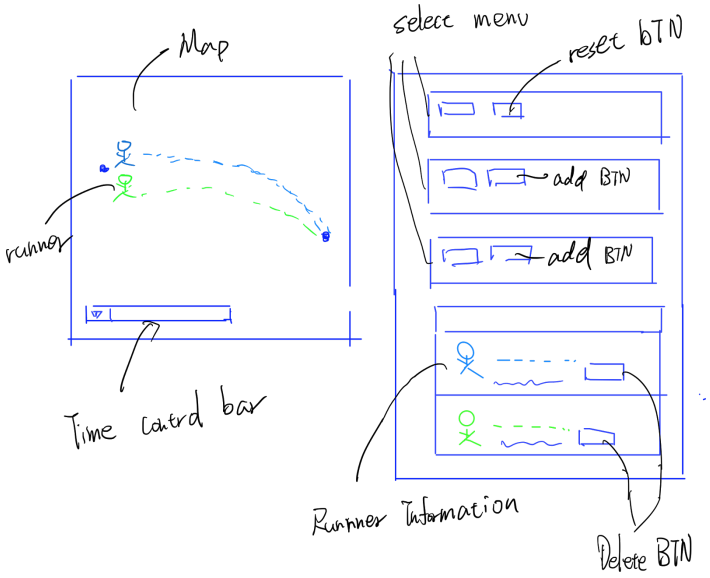


11/20

After discussion with TA, we realized that there is other group doing the same top and they want to do exactly the same visualization. Therefore, we decided to do something different and more interesting.

12/02

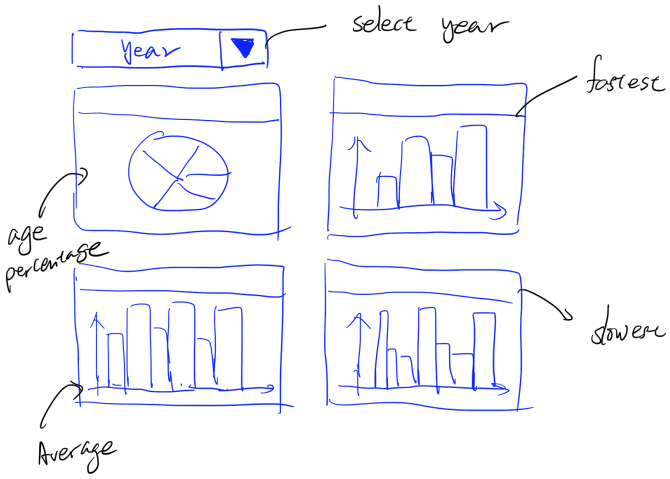
Finally, we came out some idea about showing individual records on the map and we plan to provide several filters, like year, age group, category ,gender, and each filtered result can display the fastest, the slowest and the average speed. Moreover, the people who wants to compare them record to others can search themselves by name and add on the map. The purpose of this visualization is to give the user an insight of the performance difference among others.



4. Performance Analysis

11/13

This dashboard aims to present the best, worst and average runner performance of each race group over year. Each time you move the mouse to a donut area or bar, more detailed information will be presented.



12/02

We change our year selection to a slider. This gives the user a more interactive and vivid way when using our data visualization.

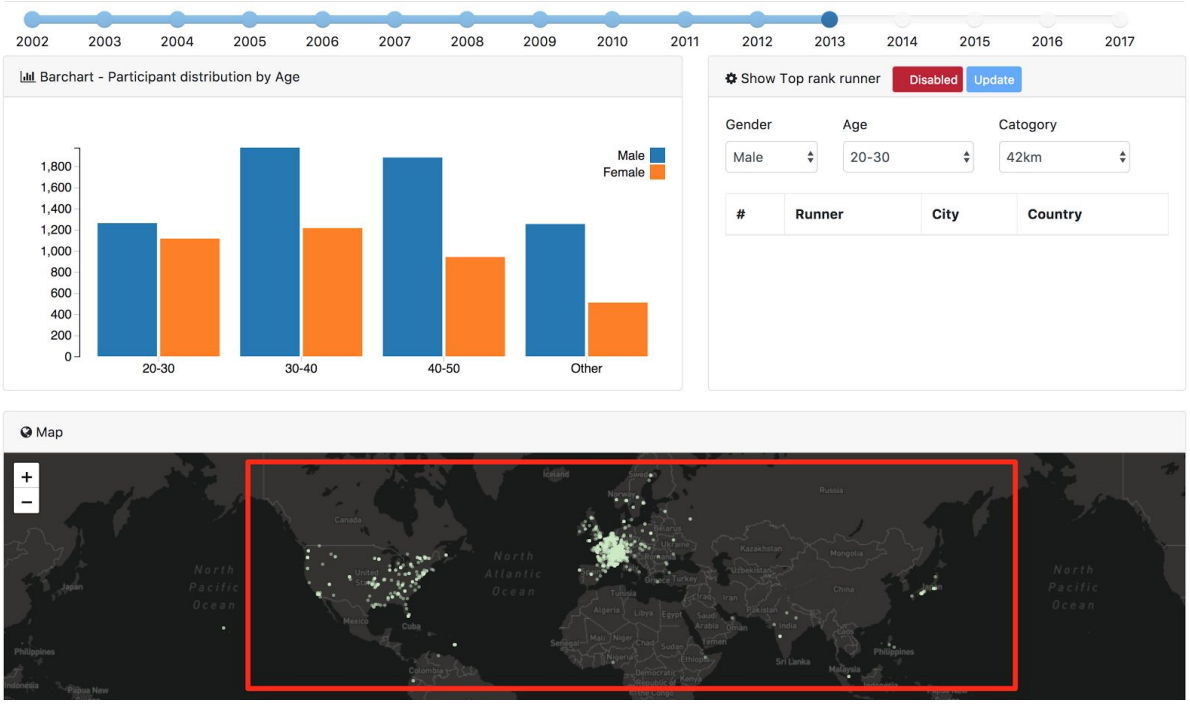
Implementation

Our datavisualizations were done by several javascript libraries:

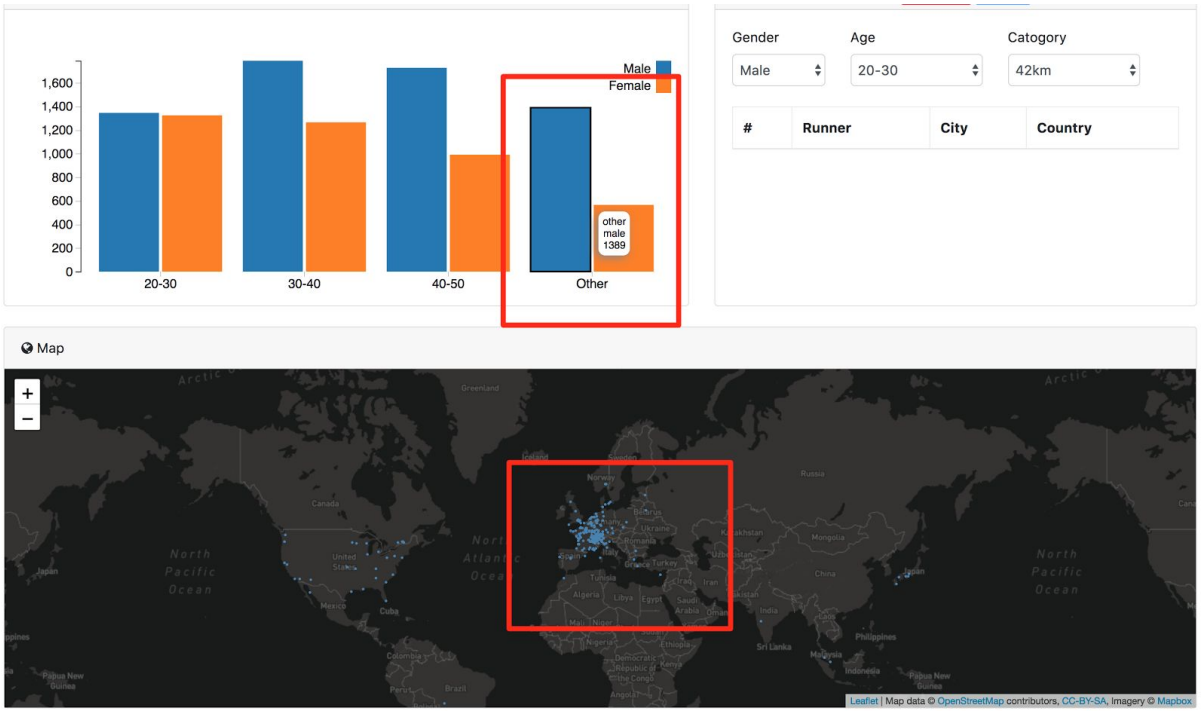
1. D3.js -
2. Crossfilter.js - Quick dataframe manipulation to filter desired data
3. Leaflet. js - provide a basic layer and control of map from MapBox
4. Leaflet TimeDimension - an time control plug-in for Leaflet.js
5. jQuery - Dom selector and some event-listener

1. Locality

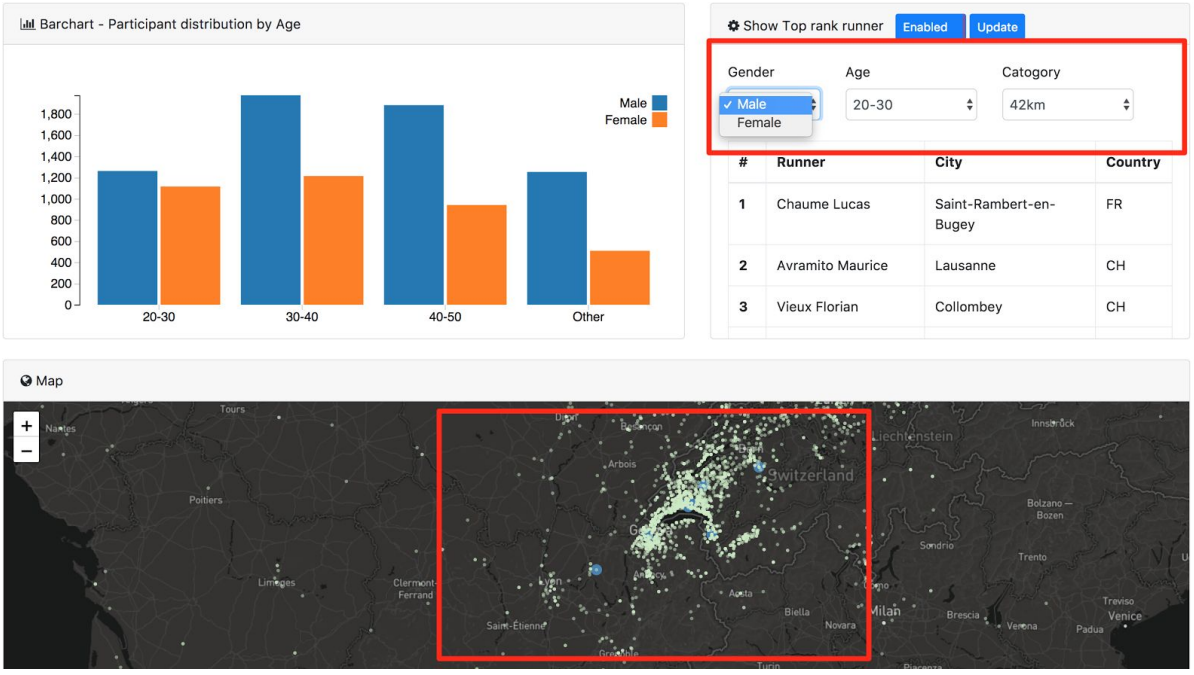
When the user manipulates the slider, the runner distribution of the selected year will be shown on the map with the semi transparent green dots. We chose a black themed map because it can contrast the green dots and it is easier to observe the distribution on the map. We used the slider as our major filter option since it provides a more interactive experience to the user compared to the basic selection drop down. Instead of using the leaflet built-in mark class, we add a svg layer using D3, which gives us more flexibility when rendering the data.



The bar chart hovering effect would display the tooltips, which illustrate the number of the group and we link the runners' home cities of this group to the map. This could be useful for people who would like to know where a certain group of people comes from.



When the user clicks the enabled button, the top 10 runners of the selected group will appear on the map with the bigger blue dot. In addition the map will re-center to the first runner. This will allow us to observe, where the best runners come from.

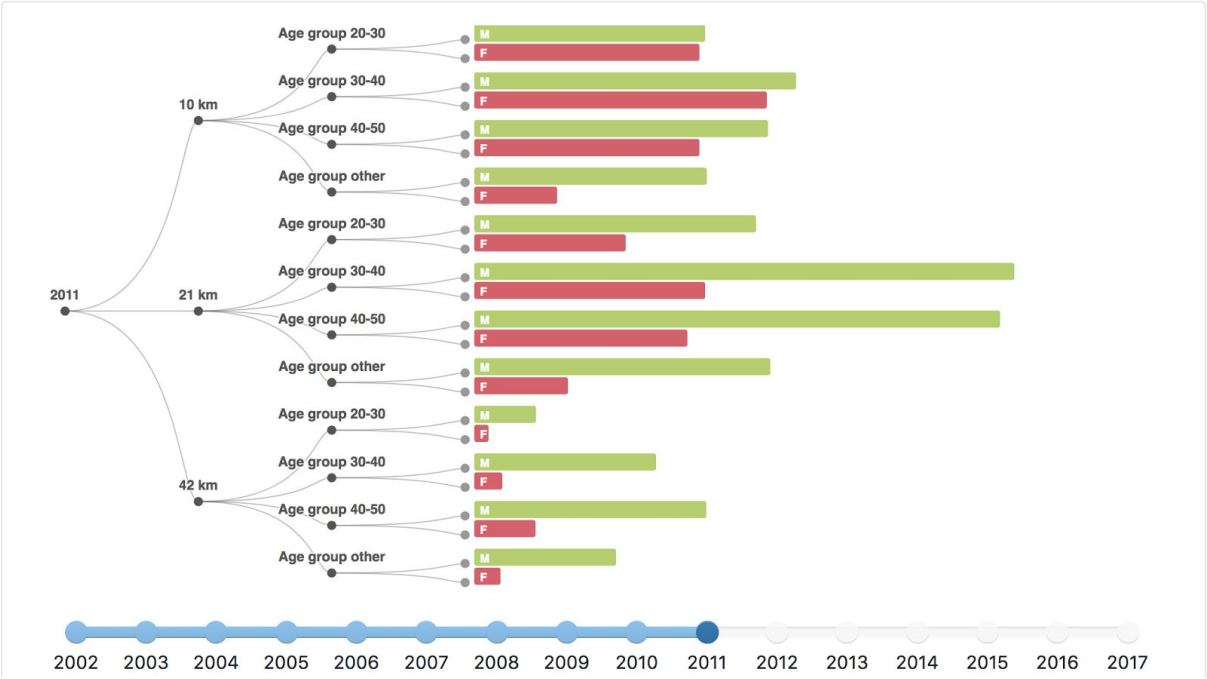


If user wants to look into a specific runner, one could simply move the mouse to the table and the map will re-center to the city where the selected runner came from, and the dot of the selected runner will change to yellow and the overall distribution will fade out for user to see the dot more clearly.

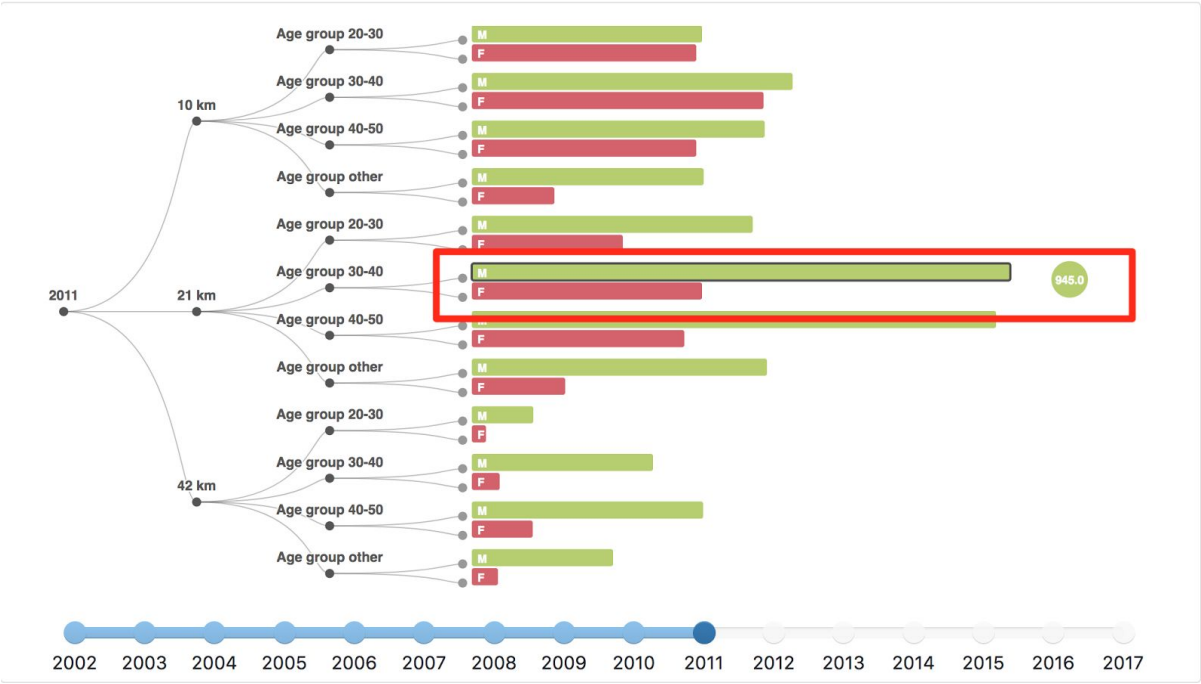
2. Race Description

Initially, the purpose of this dendrogram is to give user a brief perception about the structure of Lausanne marathon competition since the structure and the way they computed the final rank is not intuitionistic and more or less complicated. From our point of view, using dendrogram is the most efficient way to present the competition structure. Moreover, we add a histogram at the end of each last leaf in order to show the number of participants of each race, gender and age group so as to present the runner distribution of the competition. We add a time slider at the bottom of the graph so that user can observe the distribution variation over year and find some interesting tendency and insights.

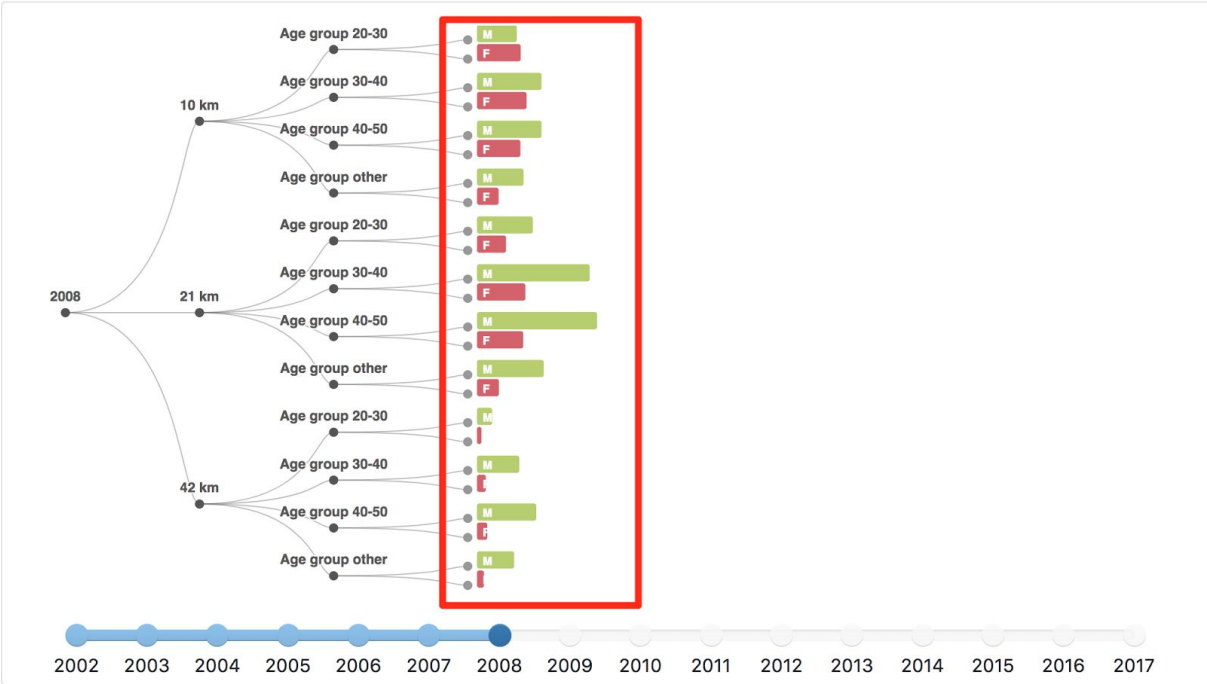
The user could easy filter and select the different year of data by sliding the slider at the bottom of this graph.



When hovering the bar, a circle tooltip will show up and display the number of people in each group. It is a neat way to show the original data rather than adding an axis on this graph.



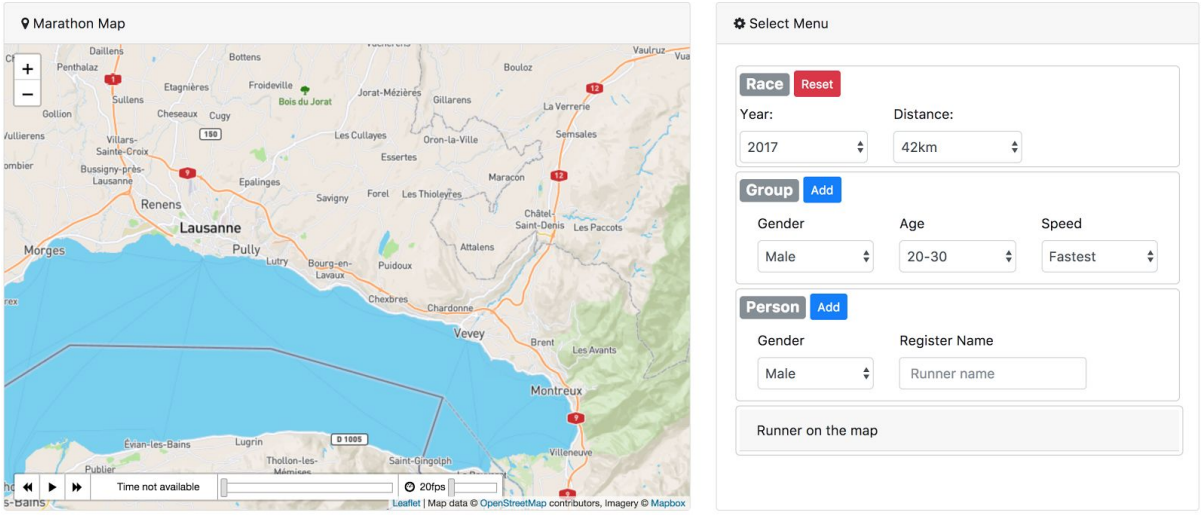
The transition effect on this graph to make it more interactive and enhance the visual effect while changing from year to year.



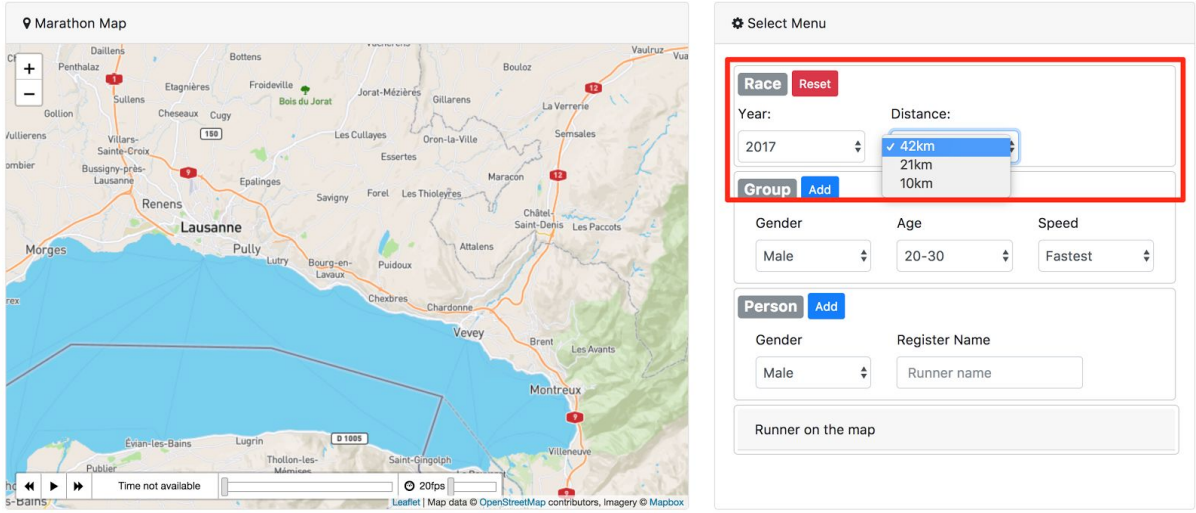
3. Speed

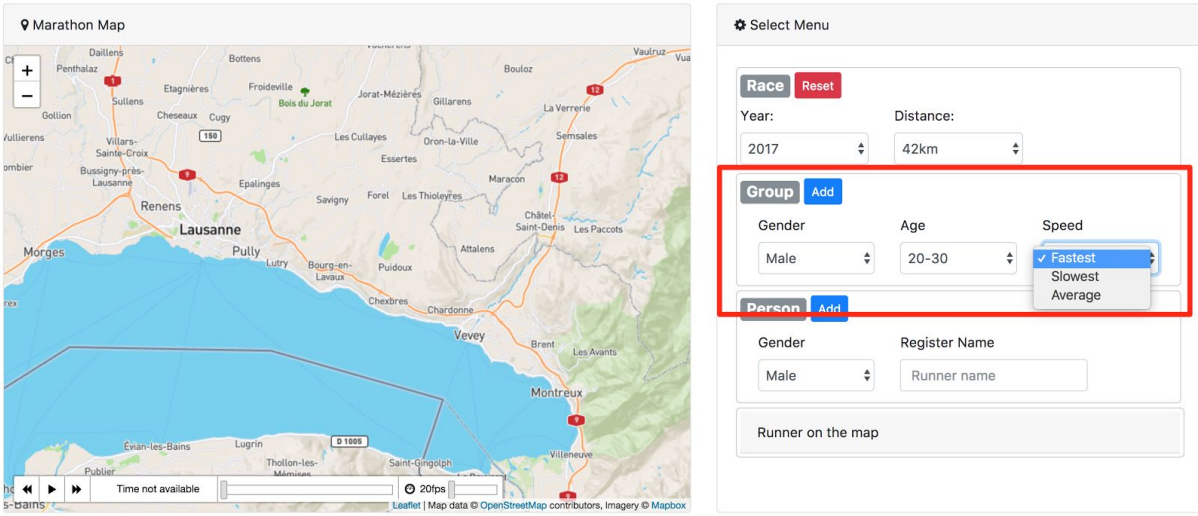
The purpose of this visualization is to visualize how the runner moves on the path along the time. And it allow the user to add more than one runners and compare their speed and show the how big difference it is among the runners. And for the runners who wants to compare their records to others can also use this visualization to illustrate the difference.

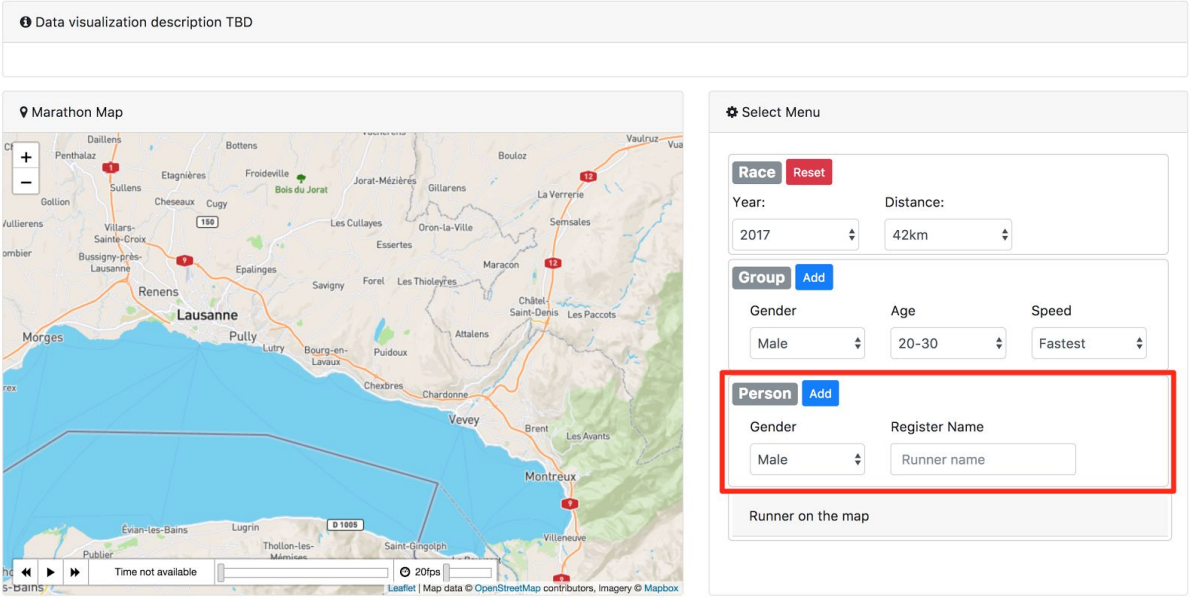
This pages contains two blocks, one is marathon map which is the route of Lausanne marathon and the other is the select menu which allows users to filter their desired player and display it on the map. Moreover, the users could select at most five people from arbitrary year, distance, group and person.



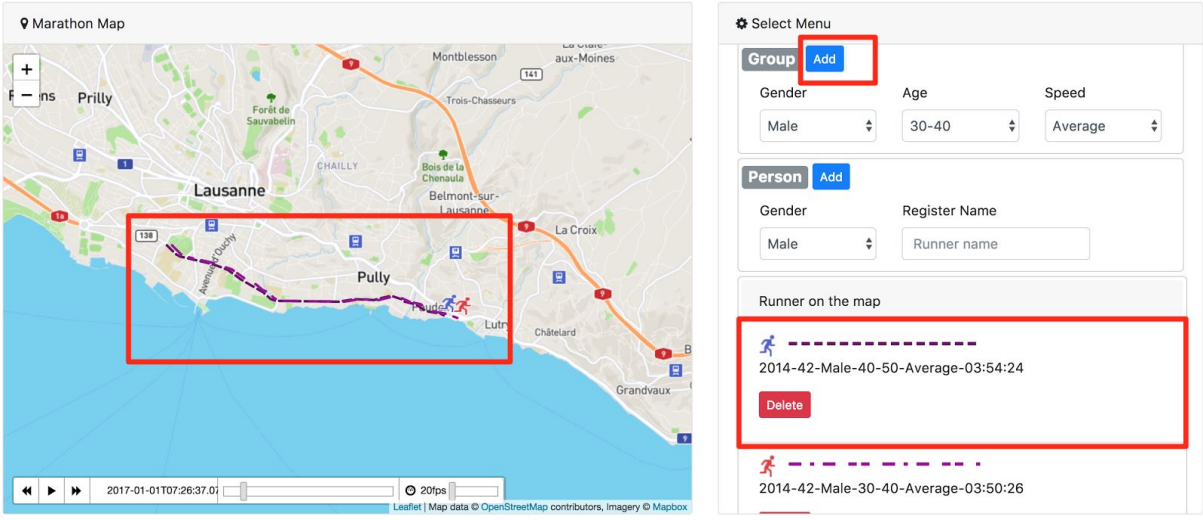
We simply use selection because it is more straight forward for user to manipulate the data. There are three section to select the desired runner. The first section is to choose the year and category of the race. And the second and third section are the options to the user, either one wants to choose a record from a certain group or to search someone by name.



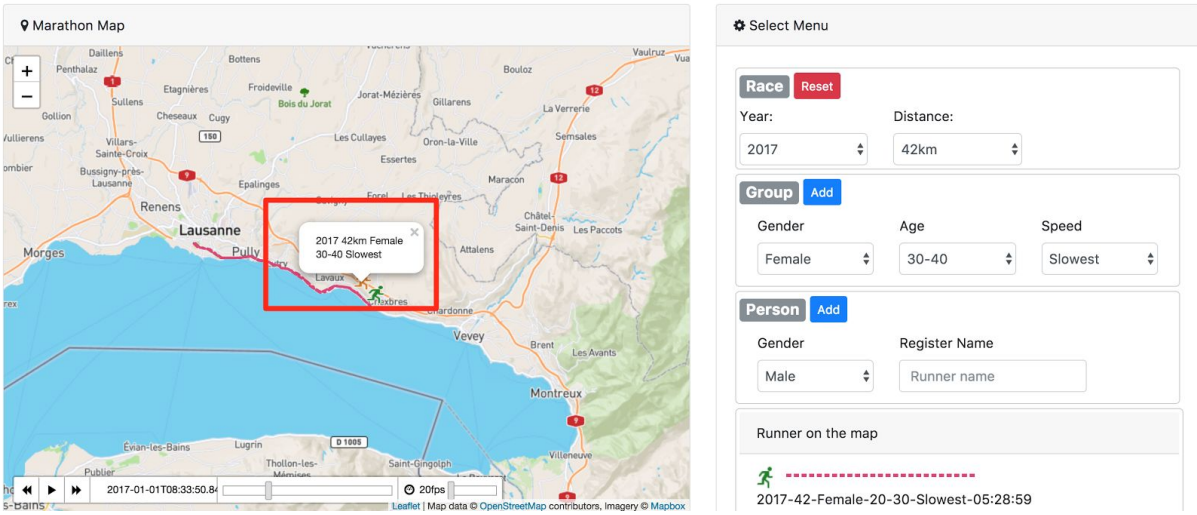




By clicking the add button, the users can add a specific runner on the map, and the information of this runner will display on the list. The position of runner is indicated by a smaller running man icon on the map and the path one had run is illustrated by the dashed line. Different runner are distinguished by different color of the icon and different dashed. This could help user to locate the runner.



Hovering over the icon on the map will display runner's information. It could also help users to identify their desired runner while they are moving on the map



4. Performance Analysis

The user could easy filter and select the different year of data by sliding the slider in the middle of this graph. And there are four charts on this page, which illustrate the performance according to the race category and gender.

When hovering to the element of each chart, it automatically links to the other three charts and highlight the corresponding value and position with the black stroke. This effect could help user to know the performance difference of the same group.

Evaluation & Conclusion

About project

In this project, we analyzed the Lausanne Marathon dataset from 4 aspects: runner locality, race structure, runner speed comparison and runner performance. From runner locality map, we observe that people from all over the world come to Lausanne to take part in this competition each year and top runners are not always from the city place even country, which implies that there are few regional factors affecting runner's performance. We shouldn't roughly conclude that runner from one country or department has better running level than others. With race structure dendrogram, we think out that as more and more people participate in 10 km and 21 km marathon races, the number of runners who join the 42 km full marathon race is slightly declining over year. Hence, we roughly claims that people become much lazier these years which made less and less people be keen on full marathon race. There exists various potential causes, for instance, working pressure, life rhythm etc. We are supposed to combine more information to draw a more rational conclusion in our future research. With the runner performance dashboard, we find that the fluctuation of average run time is not very evident, which indicates that the average runner level doesn't change a lot. Nevertheless, the best and worst runner's performance can vary dramatically. For this phenomenon, although some weather factors should be taken into consideration, the variation is quite drastic. Thus, we conjecture that perhaps those truly high level runners don't participate in this competition every year. In order to show this fluctuation in a more direct way and let user have a general perception of the runner's performance, we create the speed comparison platform. We can arbitrarily choose one or several runners in terms of their speed, gender, age, race category even name so as to simulate their running courses on the real path.

All our visualization works function well on our website. With the help of the powerful visualization library D3.js and data management library crossfilter, our graphs can not only transfer bald data into clear and intuitionistic visualization, but also animatedly interact with users according to users' diverse demands. These interactive graph provides user a more efficient and convenient way to be familiar with our dataset and check out the information they want immediately and even discover some unexpected insights. As we are three marathon enthusiasts, we plan to enlarge our dataset domain, which means that we are going to combine more relative dataset like weather dataset in order to analyze our observation in a more persuadable manner. Furthermore, we still would like to work on this project after this course, we really would like to utilize more technologies like plotly.js to enrich our website and construct more impressive data visualization for fans like us, for DataSport and for sponsors.

About course

After one semester study, we found this data visualization course is amazingly informative and instructive. The course introduces one to more parameters and considerations to account for when building data visualizations than they would probably imagine. From human perception of visuals to 2D and 3D rendering to ideal plots for different series of data, the course covers a lot of ground and introduces plenty of creative concepts.

Moreover, the project gives us a great opportunity to become familiar with D3.js in a short time and to put as much theory that we understand during the lecture as we can into practice. The project is very creative and inspiring as well, since the professor gives us great flexibility for the choice of dataset, which indicates that we can devote ourselves into a domain that we are truly interested in. In all, we really enjoy this one semester data visualization experience and have learned a lot. Of course, this is just a beginning.

Peer assessment

- ★ Preparation – were they prepared during team meetings?
  Yes.
- ★ Contribution – did they contribute productively to the team discussion and work?
  Yes.
- ★ Respect for others' ideas – did they encourage others to contribute their ideas?
  Yes.
- ★ Flexibility – were they flexible when disagreements occurred?
  Yes.