ETL Project Write-Up

Arthur Edwards, Terisha Kolencherry, Pooja Nagrecha, Jack Tambert

Extract

This data analysis looks at data from two triathlon organizations, USA Triathlon (USAT) and National Senior Games Associations (NSGA). A senior triathlete is defined by the NSGA as someone 50+ years of age. While the layout of the data is different between the two sites, both contain information about total time taken to complete the triathlon, the sex of the triathlete, age data of the triathlete, and state data.

For both pages we used splinter to open up a browser and go to the respective triathlon results main page.

National Senior Games Association

For the NSGA data the main site was divided into two tables - one male and one female and the anchor tag text had the age information for all competitors in the ensuing results table. We set up three empty lists - sex, age, and links. We then set a variable for sex equal to "Male" since the first table was all males. We used a for loop to go through each of the tables to find the rows and cells within the table and bring back the age and link information from the anchor tags. We put this information into a dataframe and used string concatenation to generate full URLs for each results page.

We then used another for loop in combination with Pandas' to_html function to bring back the tables on each results page and put it all in one dataframe along with the age and sex information. A snapshot of the raw data is presented below:

	Unnamed: 0	Name	State	S	T1	B	T2	R	Time	sex	age
0	1st	REMBAC, Ross	AZ	00:06:29.000	00:04:58.000	00:35:22.000	00:00:47.000	00:23:43.000	01:11:19.000	Male	50-54
1	2nd	GALLARDO, Paul	NM	00:08:07.000	00:03:44.000	00:38:18.000	00:01:09.000	00:20:24.000	01:11:42.000	Male	50-54
2	3rd	KASSA, JOE	NM	00:09:39.000	00:04:28.000	00:38:55.000	00:00:52.000	00:24:36.000	01:18:30.000	Male	50-54
3	4th	HOBBS, Bob	KY	00:11:35.000	00:04:06.000	00:39:53.000	00:00:46.000	00:24:30.000	01:20:50.000	Male	50-54
4	5th	WYATT, James	TX	00:07:23.000	00:05:32.000	00:42:55.000	00:01:21.000	00:27:36.000	01:24:47.000	Male	50-54

USA Triathlon Data

The USA Triathlon data was not able to be scraped by means of pd.to_html, because each table is structured as cells within the page, rather than an html table. Also, we needed to take use of Splinter to remotely control the web pages. We then obtained each URL leading to the page for each race. Utilizing the URLs, we looped through each of the URLs to grab the data from each page. To grab the data from each page, we searched for the

Transform

Since the data is presented differently on each page, our group had to clean each dataset to provide commonality.

National Senior Games Association

After scraping the data from the website, we first split the data in the Name column into First and Last Names and appended that data to the original dataframe. Then we dropped all rows where the time was either "-",

indicating a no-show, or "DNF, indicating that the participant didn't finish. Afterwards, we ranked and reindexed the data based on time from fastest to slowest.

Rank	First	Last	State	Time	gender	age
1	Dave	CAMPBELL	CA	01:09:09.000	Male	60-64
2	Derrill	STEPP	CA	01:10:31.000	Male	55-59
3	Ross	REMBAC	AZ	01:11:19.000	Male	50-54
4	Vanessa	COOK	NV	01:11:54.000	Female	50-54
5	Louis	SALAZAR	NM	01:12:30.000	Male	55-59
...	...	...	...	...	...	...
102	ERNEST	SCHILLINGER	VA	02:41:56.000	Male	85-89
103	Kathleen	TILLER	OH	02:43:18.000	Female	70-74
104	Linda	PLEIN	NM	02:45:34.000	Female	75-79
105	Patricia	STOLTENBERG	IL	02:51:09.000	Female	60-64
106	Patsy	LILLEHEI	MN	02:54:41.000	Female	75-79

USA Triathlon Data

After scraping the data from the website, we binned the data into age groups using increments similar to the ones in the NSGA. One potential issue that was considered was the difference in triathlon types. The NSGA Triathlon is a sprint triathlon (750 m swim + 20 km bike + 5 km run) The data from USAT has multiple different lengths (Sprint, Olympic, Half-Triathlon). However, sprint triathlons are the default view in the race results pages so the data pulled from USA Triathlon is comparable.

Rank	First_Name	Last_Name	Sex	Age	State	Time	Race	Race
1	Michael	Alexander	M	50-54	FL	01:02:29.000	87.912	Bartow Blarney Triathlon
2	Mark	Hulbert	M	50-54	FL	01:04:06.000	85.695	Bartow Blarney Triathlon
3	Brian	Durden	M	35-39	FL	01:05:30.000	83.863	Bartow Blarney Triathlon
4	Rodney	Carter	M	40-44	FL	01:05:54.000	83.354	Bartow Blarney Triathlon
5	James	Hooppaw	M	25-29	FL	01:06:10.000	83.018	Bartow Blarney Triathlon
...	...	...	...	...	...	...	...	...
16	Jason	Blackman	M	35-39	FL	01:23:43.745	77.732	HITS Triathlon Series: Sarasota, FL
17	Katie	Hammond	F	35-39	FL	01:24:08.040	85.094	HITS Triathlon Series: Sarasota, FL
18	Mateus	Arruda	M	30-34	FL	01:24:25.790	77.087	HITS Triathlon Series: Sarasota, FL
19	Hunter	Carey	M	20-24	IA	01:25:26.741	76.17	HITS Triathlon Series: Sarasota, FL
20	Doerte	Fehseehlert	F	50-54	0	01:25:30.953	83.719	HITS Triathlon Series: Sarasota, FL

Load

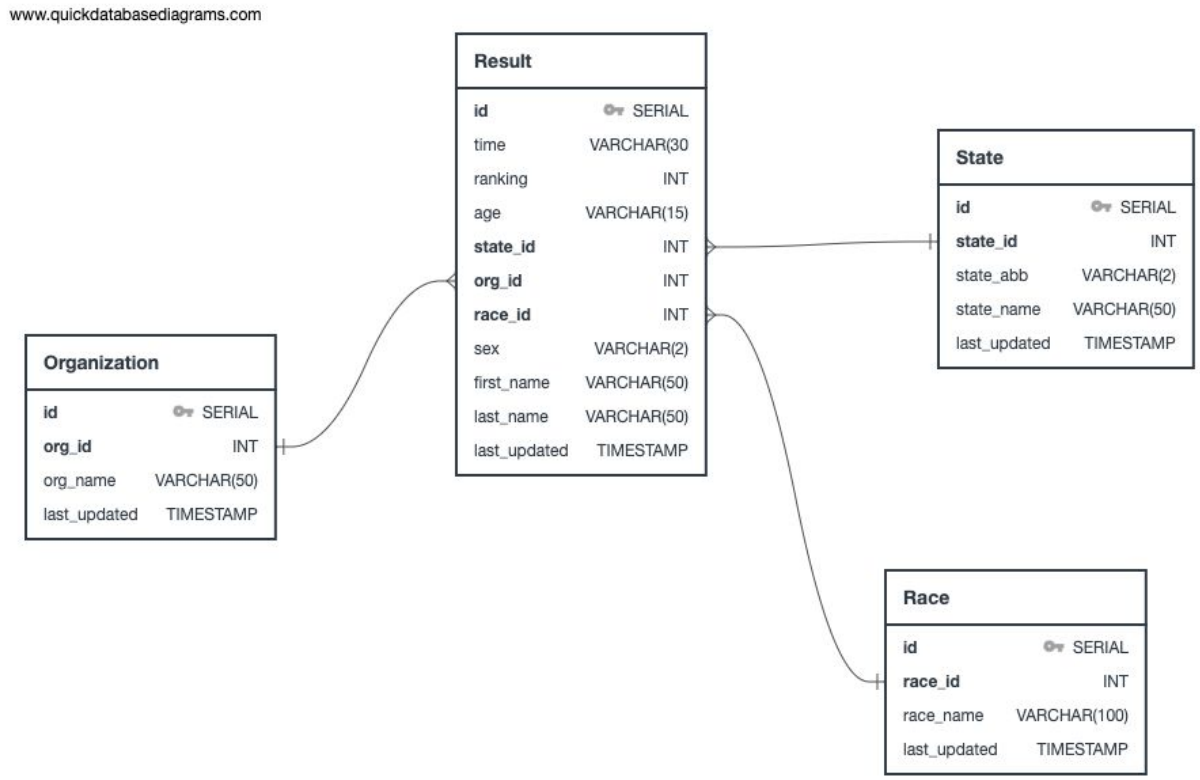
Our database consists of four tables - three lookup tables for state data, organization data, and race metadata plus a table for race results. Below is the layout for the table:

Race Table - this table is a lookup table containing the race metadata including the name and race of the date. Primary key is race_id, a serial integer automatically assigned by Postgres . Last updated will also be assigned by Postgres as the datetime of the last time the data was updated.

Organization - this table is a lookup table which will identify the source of the race results. Primary key is organization_id, a serial integer automatically assigned by Postgres. Last updated will also be assigned by Postgres as the datetime of the last time the data was updated.

State_Data - this table is a lookup table which will identify the State that the triathlete is from. Primary key is state_id, a serial integer automatically assigned by Postgres. Last updated will also be assigned by Postgres as the datetime of the last time the data was updated.

Results - this is our main table which has information on the first name, last name, age range, gender, race time, and ranking. The table will have foreign keys for race_id, organization_id, and state_id.



Limitations of the Dataset and Recommendations

One limitation of the selected datasets is the difference in sample sizes between USA Triathlon and NSGA. NSGA has one race versus the multiple different races that USA Triathlon runs. Additionally, USA Triathlon races have wider participation in each race since the races aren't for specific age ranges and athletes don't have to qualify for the races. If this dataset were to be used for statistical analysis - for example, using a two sample t-test to see if there's a statistically significant difference in the average race time for USA Triathlon athletes vs NSGA athletes - we would expect to see a larger spread of data for USAT.