Project: Tour de France

1 Le Tour 2025

1.1 Tour de France

The *Tour de France* is an annual professional road cycling competition held primarily in France, with occasional routes passing through neighbouring countries. The competition spans three weeks in July, covering a distance exceeding 2,000 miles (3,500 kilometres). The competition consists of successive stages. Each stage may be an individual time trial or a mass-start stage where all riders commence simultaneously. The ultimate victor of the Tour de France is the cyclist who completes all stages in the shortest cumulative time.

The competition features multiple ranking systems called "classifications" that order riders and teams based on specific performance criteria. Among these, the General Classification determines the overall winner. Additional classifications are the Points Classification for sprinters, the King of the Mountains Classification for climbers, and the Young Rider Classification for participants below a specified age threshold.

The Tour de France features a selection of professional cycling teams participating in the competition. These teams are typically composed of riders who work together throughout the competition to support each other's efforts. The Tour de France includes a mix of WorldTour teams (the highest level of professional cycling) and wildcard invitations. The teams come from various countries, although many are based in Europe. Some of the most prominent cycling nations, including France, Italy, Belgium, Spain, and the Netherlands, often have multiple teams in the Tour de France.

A Tour de France cycling team typically consists of 8 riders. The riders in a team may come from various countries different from the country of the team. Each team is allowed to field a squad of eight riders who compete in all the stages of the competition. These riders work together throughout the competition to support each other, with different members having specialised roles, such as sprinters, climbers, and domestiques (supporting riders). It is worth noting that each team may have a designated leader who is the primary contender for the General Classification and wears the Yellow Jersey if they lead in the General Classification. The other team members will work to support the leader's efforts.

1.2 Application Specification

1.2.1 Teams

Each team is uniquely identified by its name and belongs to exactly one country. The database must allow countries to be recorded even if no team is associated with them. The database does not have to include all countries worldwide, but it should include all countries involved in Tour de France 2025. Involvement can be based on riders and teams as well as start and finish locations.

A country is uniquely identified by the IOC code [1]. This is a 3-letter code. For instance, France is identified as `'FRA'` using this code. We also record the name of the country which should also be unique. Lastly, we record the region of the country. For instance, France is in Europe.

A team is composed of one or more riders with a rider belonging to exactly one team. Unlike country, we do not want to record the team with no riders. A rider is uniquely identified by their bib number. We also record the name and date of birth of the rider.

A rider belongs to at most one country. It is possible that we have a rider with no country data. Similar to before, we want to record the countries with no riders. A rider need not have the same country information as the team. For example, we may have a French team with a German rider.

1.2.2 Race

Tour de France race is completed in stages. In each stage, teams race from one location to another. A location belongs to exactly one country and a location is identified by its name. We only record locations that are used in at least one of the stages as origin, destination, or both.

Each stage is held in a single day. There are five types of stages. Three of them depend on the types of the environment, namely, **flat**, **hilly**, and **mountain**. Two stages are time-trials called **individual time-trial** and **team time-trial**. We record both the type of the stage as well as the length of the stage given in kilometres (i.e., the distance between the two locations).

Individual results for each rider are recorded for each stage. We record the total time as the number of seconds. Since there can be different riders that can finish at the same time, we also record the rank that depends on the order in which the rider crosses the finish line. Additionally, time bonuses are awarded at the end of each stage for the first few riders to cross the finish line. Potentially, penalties may also be incurred during a stage. The total time for each rider can be calculated by adding the time it takes to finish, subtract the time bonus, and add the time penalty.

Ideally, there should not be two different riders with the same rank for the same stage. Also, there should not be "gaps" in the rank. For instance, if there is a rider of rank 1 and rank 3, there should also be another rider with rank 2. The best rank is rank 1.

1.2.3 Exit

In some cases, a rider may exit from the race at the beginning of a particular stage. We record the reason for the exit. If a rider exits at some stage S, there should not be any individual result recorded for the rider starting from stage S onwards.

This allows us to not insert the individual result of a rider that has not exited yet (e.g., maybe due to the data not yet available) but it prevents us from inserting an individual result of a

rider that has exited. For now, there are only two reasons, namely "withdrawal" and "DNS" (i.e., "do not start"). However, there may be other reasons to be added in the future.

1.2.4 Rest Day

Certain days are designated as "rest days". There is no race (i.e., no stage) during a rest day. A rest day should only last for one day. In other words, there should not be a consecutive rest days. Additionally, there should only be two non-consecutive rest days for the entire competition. For simplicity, you may assume that there are no designated "rest days" before the earliest stage and after the latest stage currently in the database.

For instance, this allows us to insert the stages in the order shown on the left but not in the order shown on the right.

- Day 1
- Day 3 (*assume Day 2 is first rest day*)
- Day 5 (*assume Day 4 is second rest day*)
- Day 2 (*Day 2 is no longer rest day*)
- Day 4 (*Day 4 is no longer rest day*)

- Day 1
- Day 4 (*2 days of rest*)

- Day 1
- Day 3
- Day 5
- Day 7 (*3 rest days*)

1.3 Awards

We can find the current leader after each stage based on the **accumulated adjusted time** each rider takes to complete the stage after computing the bonus and penalties as specified above (i.e., time - bonus + penalty). The rider who completes all stages in the shortest cumulative adjusted time is declared as the ultimate victor! There is also an award for the best team. The team with the **lowest aggregate time** (i.e., the lowest sum of the three best riders' times) is awarded the "Best Team" classification. You may find other potential results from the website [2].

2 Tasks

Your company, **Apasaja Private Limited**, has been commissioned by the team EF Education - EasyPost to analyze their poor performance in Tour de France 2025.

One of their interns managed to scrape some raw data from the Tour de France website. Unfortunately, as the intern is not well-versed in database design, the data is given as a single file in a *comma separated value* format (i.e., `csv`). There is a total of 26 columns in the file `tdf-2025.csv`.

Basic Information

day	stage	bib	rank	time	bonus	penalty

Starting Location Information

start location	start country code	start country name	start region

Finish Location Information

finish location	finish country code	finish country name	finish region

Additional Information

length	type	rider	team	dob

Rider Country Information

rider country code	rider country name	rider region

Team Country Information

team country code	team country name	team region

You are also given another set of data obtained by a scout from EF Education - EasyPost. This data contains information about the riders who exit the competition according to §1.2.3. There is a total of 3 columns in the file `tdf-exits.csv`.

Rider Exit Information

bib	stage	reason

2.1 Roadmap

To analyze the performance of EF Education - EasyPost, you proposed the following roadmap. This will be your roadmap for the project.

1. Provide a **minimum viable product** (MVP) by showing the advantage of using database over `csv`.
   - Create a table and insert some data from `csv` to the database following your table.
2. Improve the MVP by applying entity-relationship model.
   - Provide the entity-relationship diagram, translate into schema, insert all the data from `csv`, and answer basic question about the competition.
3. Enforce all the constraints using triggers.
   - Provide the stored procedures and triggers given a schema.

We will provide more details about the individual tasks in the roadmap as the semester progresses. It is possible that parts of the data contains *inconsistencies* as the intern may not be an expert at web scraping. If you design your database correctly, this should be easily captured.

If there are inconsistencies, you need to make sure that it is indeed an actual inconsistencies. Once confirmed, you may need to update the data to resolve the inconsistencies manually. Our description of Tour de France may be an *idealized* description. As such, you should follow our description as you are not expected to be an expert in Tour de France.

For simplicity, we assume there are only `INSERT` and `UPDATE` operations on our tables. In other words, once you have inserted a rider R, you are not allowed to remove them. However, you may modify them to any other valid values.

1. (**P01**) **Creating and Populating Tables with Constraints.**

   Follow the instruction on Canvas Assignments > P01: Tables for submission information.

   (a) **Constructing Tables.**

   Construct tables usign SQL data definition language (DDL) consisting only of `CREATE TABLE` with the following five basic integrity constraints: `PRIMARY KEY`, `UNIQUE`, `NOT NULL`, `FOREIGN KEY`, and `CHECK`. You should not have any other statements including **but not limited to** `CREATE DATABASE`, `CREATE SCHEMA`, `CREATE VIEW`, `CREATE ENUM`, `CREATE TYPE`, `CREATE FUNCTION`, `CREATE PROCEDURE`, `CREATE TRIGGER`, etc. You should use the default schema (i.e., do not alter the `search_path`, etc) as well as any other setting. However, you are allowed to use `ALTER TABLE` if necessary.

   Ensure that all the data can be accommodated (i.e., you have sufficient number of columns, the data type is big enough to handle the data, etc). Avoid catering for unnecessary, inconsistent, or redundant data. Propagate updates to maintain referential integrity when needed. Only propagate deletion where it makes sense.

   Note that you do not have to replicate the `csv` columns structure exactly. In fact, you are encouraged to transform the data (e.g., splitting columns, adding columns, removing columns, etc) to enforce constraints using the five basic integrity constraints above. However, you should not add artificial keys when there are natural keys available. This includes **but not limited to** the use of `SERIAL` data type.

   You should only use **atomic data**. Note that atomic data depends on the *semantics* and not only on type. For instance, `TEXT` may be atomic (e.g., first name) or non-atomic (e.g., concatenation of all phone numbers) depending on usage.

   Note that **you are not able to enforce all constraints** and **you are not required to enforce all constraints**. You are only required to enforce as many constraints as possible using the five basic integrity constraints above.

   Write your `CREATE TABLE` statements in the schema file named `P01-schema.sql`.

   (b) **Inserting Stage 1 Data.**

   Write the `INSERT INTO` statement to insert the data for Stage 1. Stage 1 of Tour de France 2025 was on 4 July 2025 from Barcelona to Barcelona. The insertion should follow the tables you created above. Since there may be foreign key constraint on your table, you may need to insert to multiple tables. Ensure that the execution of both table creation and data insertion have no error. There should not be any inconsistencies in the data for stage 1.

   You should not have any other statements besides `INSERT INTO` for this part. This includes **but not limited to** creation of *staging tables* as well as the use of `\copy`. Any statements other than `INSERT INTO` statements will be penalized heavily.

   Write your `INSERT INTO` statements in the data file named `P01-data.sql`.

References

[1] *International Country Codes*. `https://www.worlddata.info/countrycodes.php`. [Online; last access January 2026].

[2] *Tour de France 2025*. `https://franceletour.com/`. [Online; last access January 2026].