Jordan Lee (30002218)

Data Approval

https://www.kaggle.com/ralle360/historic-tour-de-france-dataset

The Tour de France is the world's most prestigious and famous bicycle race, consisting of 21 race stages over three weeks. Although it is the most watched sporting event in the world, it can be difficult for a new spectator to understand due to the events focusing on different disciplines and, unlike many other sports, the changing race course from year to year. By visualizing the historical data of the Tour de France, it can provide a greater appreciation for the sport and how it has evolved over time.

I will be using a dataset found on www.kaggle.com, generated by RasmusFiskerBang. It contains race stage information from the first Tour in 1903 up to the 2017 Tour de France. The information was gathered from Wikipedia, and there are 2336 rows and 8 columns.

Each row represents a single stage in a given year of the Tour de France; the columns provide details about that stage, including information such as:

- **Stage** - The stage number in the Tour of that year
- **Date** - The date that the stage occurred on
- **Distance** - The number of kilometres covered by the stage
- **Origin** - The name of the city where the stage starts in
- **Destination** - The name of the city where the stage ends in
- **Type** - The main type of the stage (eg. flat, time trial, mountain, etc.)
- **Winner** - The name of the winner of the stage
- **Winner_Country** - The country of origin of the stage winner

The dataset contains a variety of types of data, primarily nominal but some quantitative as well. One notable missing metric is the time of each winner, which could have been used to calculate things like pace. Also, the dataset only includes up to the 2017 tour, but the rows missing can easily be added manually due to the small size. The source of the data is from Wikipedia and therefore may not be as accurate as data from the official Tour de France website, but the convenience of being in `.csv` format allows it to be more easily compared and cross-referenced with similar datasets for errors, in addition to creating visualizations. The large number of rows can help identify outliers if there are any minor inaccuracies, and is therefore appropriate to be used for viewing the general trends.

I will be using the entire dataset `stages_TDF.csv`, as there are many interesting ways to represent the different types of information. For example, the average timeframe that each Tour occurs within the year displayed over a calendar, or the most common cities to host stage starts/ends and their most recent year that they hosted shown on a map. As an avid cyclist who found the Tour de France confusing to follow at first, I hope that creating interactive visualizations using this data can help demystify and provide insight on the history of the Tour.