

Project Proposal II

Updates to your proposal – what has changed in terms of websites scraped, variables acquired and generated, questions addressed, visualizations planned, and format of your final product

Websites Scraped:

Tour De France Winners https://en.wikipedia.org/wiki/List_of_Tour_de_France_general_classification_winners

Tour De France Stage Wins per Rider https://en.wikipedia.org/wiki/Tour_de_France_records_and_statistics#Stage_wins_per_rider

Athletes in the top 10 of any year after 1998 and whether they have been caught doping as well as a general summary table. https://en.wikipedia.org/wiki/Doping_at_the_Tour_de_France

Variables

Tour De France Winners: Information on the winner of any given year including name, country, year, sponsor, distance, time, margin won my, stages won, stages in lead.

Tour De France Stage Wins per Rider: This table gives a list of names, all time rank (Based on stage wins), and number of total stage wins.

Athletes in the top 10 of any year after 1998 and whether they have been caught doping as well as a general summary table

We have information on the top finishers of the Tour de France from 1998 to 2015. This information includes finish place, rider name, team, time/margin, and notes. Notes will provide ample information after string analysis of the doping status of the rider.

An addition table also provides doping status of Tour winners since 1961. This second table includes year, name, status, and detail of the doping history.

New Variables / Variables generated

Not much has changed on our plans for variables generated, however we have found more information on doping which we can merge into our original data sets. We will be generating many variables regarding athlete doping status. We will do this through intensive string analysis. We have also gained information on whether a person was stripped of a title or not.

Questions Addressed

New questions we are able to address are:

How has doping disqualifications have changed the top 10 finishers of Tour in a given year? How many more athletes confess to doping vs testing positive?

In general we are better set to answer the questions mentioned in the original preliminary report.

An update on how your data scraping is coming; ideally you will have already put together your primary data set(s) by Dec 4

Well. We have scraped all data and plan to have relevant data merged before class on Tuesday.

Names of files where I can find your R code

Scraping&Cleaning.Rmd (In both of our submit folders and on GitHub).