

# Predict River Length

Team : Technopreneurs

**Task:** Estimate rivers by length

**Description:** Given a set of rivers, estimate their length

**Detailed Description shown to participants:** "We will now ask you a series of questions. For each question, we will show you a River Name, and ask you to estimate its length. You will have to fill your estimate in the box provided. Please note that the length should be given in kilometre (km)"

**Input Type:** Text

**Corpus:**

Wikipedia's "List of rivers by length"

[http://en.wikipedia.org/wiki/List\\_of\\_rivers\\_by\\_length](http://en.wikipedia.org/wiki/List_of_rivers_by_length)

**Representative tasks methodology:**

Sample 20 rivers from this list using "Simple Random Sampling".

**Justification:** Since the size of the dataset/corpus is very small (176 rivers), Simple Random Sampling is sufficient to generate a set of representative tasks.

**Answer type:** Point estimation.

**Code:** (To generate set of representative tasks)

```
#Language: R

#Read data from csv file
mydata = read.csv("river_list.csv")

#Generate a random sample of size 20 from the data
mysample <- mydata[sample(1:nrow(mydata), 20, replace=FALSE) ,]
```

# Predict Number of Goals

Team : Technopreneurs

**Task:** Goals Prediction in a Match

**Description:** Given the league name, league ranking of teams before the match, predict the number of goals that will be scored during the match \*

**Detailed Description shown to participants:** "We will now ask you a series of questions. For each question, we will show the details of a soccer match from MLS Regular Season, and ask you to predict the number of goals that will be scored in that match. You will have to fill your prediction in the box provided. Please note that you can find the current league standings here: <http://www.mlssoccer.com/standings/2015>"

**Input Type:** Text (Tabular Form)

**Corpus:**

MLS Schedule for April 2015 <sup>†</sup> (csv format)

<http://goo.gl/u9qQyI>

**Representative tasks methodology:**

Sample 20 matches from this list using "Weighted Sampling without replacement". The weights assigned to matches that are closer to the survey date will be higher so that prediction task becomes easier.

**Justification:** Weighted Sampling where weights are more for matches nearer to current date is used to generate a set of representative tasks. This approach will also eliminate the chances of random guessing for later matches. The weighted sampling is based on <http://epubs.siam.org/doi/abs/10.1137/0209009>

**Answer type:** Point estimation.

Note: Please find code on next page.

---

\*Task description was changed after the following comment by Dr. Sharad Goel on Piazza: "I think it would be interesting to ask about \*future\* games, rather than games that have already happened. For example, you could ask about MLS games that are schedule to happen in April."

<sup>†</sup>Official Schedule: <http://goo.gl/kVjV8x> — Complete csv: <http://goo.gl/xykM7x>

**Code:** (To generate the set of representative tasks)

```
#Language: R

weighted_Random_Sample <- function(
  .data,
  .weights,
  .n
){
  #This section of code is slightly incorrect,
  #Our team is currently working on it.
  key <- runif(length(.data)) ^ (1 / as.double(.weights))
  return (.data)
  # Currently returns all rows,
  # This line will be changed in the final version
}

#Read data from csv file
mydata = read.csv("MLS_Schedule_April_2015.csv")
myData = read.csv("MLS_Schedule_April_2015.csv")

mysample <- mydata[sample(1:nrow(mydata), 20, replace=FALSE) ,]

#Manipulation with date field
mydata$Composite <- strptime(mydata$Composite, "%Y-%m-%d_%H:%M")
dateFields <- as.Date(mydata$Composite, "%Y-%m-%d_%H:%M")
N <- length(dateFields)
diff <- dateFields[1:N] - Sys.Date()
weights <- as.list(diff)
weight <- t(weights)

#Final call
weighted_Random_Sample(myData, weight, 20)
```