

Creating interactive modules for teaching with sports data

Nick Clark, Robin Lock, Ivan Ramler, Michael Schuckers

USCOTS July 2025

Sports Content for Outreach, Research, and Education



Why Sports?



Data availability

Lots of **public data**,
freely available to anyone

Wide **variety** of data,
problems, and methods

Popularity

Sports are widely
popular. In 2019, 154.4
million U.S. viewers
watched live sports at
least once per month

Many students start as
subject-matter experts.

Real-life **validation**

Transferability

Problems are **analogous**
to those in non-sports
applications.

Experience in sports
translates to other fields

Sports are a **controlled**
environment. Potentially
easier to start with

Examples of Sports Analytics Problems

Teams

- Player personnel decisions, evaluate player performance (trades, FA signings)
 - How much is a player contributing to his/her team in terms of goals/points/runs?
 - How much is a player worth on the open market in terms of salary cap dollars?
 - How much is a player worth to us?
- Coaching decisions
 - Should we bunt? go for it on 4th down?
 - When should we pull our goalie?
 - Given locations of the players and ball/puck, what do we expect to happen? What is the best decision?
- Draft decisions
 - Who should we draft with our next pick?
 - Who will be available to draft?
 - How much is our draft pick worth?

Media - Talk about those decisions, team ratings and playoff probabilities

Betting - Many of the above problems, or variants of them

League - What realignment would minimize travel? What schedule would max. fairness and min. travel?

Business - Predicting demand for a game based on day, month, opponent, etc

Learning Objectives

Most of those can be answered reasonably with undergraduate level statistics and data science tools.

Students get experience with

- Solving real problems
- Joining data from multiple sources, working with several different types of data
- Data exploration/visualization
- Multivariable thinking, need for regression or something else
- Modeling
- Interpretation
- Etc

Most of the data is publicly available, or can be done with public available alternatives

Exception: player tracking data

Sports Analytics in Education

Dozens/hundreds of educators around the world have developed

- hundreds of examples,
- using multiple sports, and
- focusing on a variety of statistics and data science topics

Missing:

- Standardization
- Completeness
- Consolidation/centralization
 - Content creation is decentralized = good
 - Content is decentralized = less convenient
- Industry/media perspective

Enter SCORE!



SCORE with Data: building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

Unique NSF-funded project combining academia, professional sports, and media to build a repository of educational materials for statistics & data science via sports applications/analytics

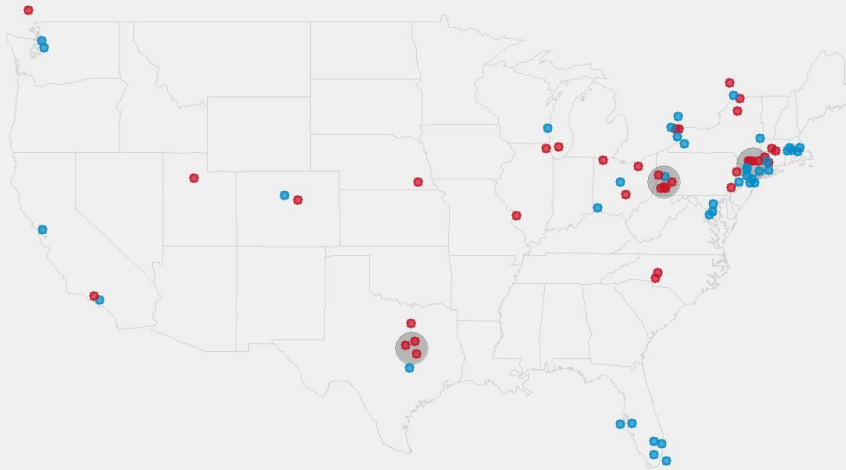
Strong emphasis on outreach, inclusiveness, and building pipelines

Sustainable National Network



Map of Initial Hubs and Confirmed Partners

• Academic Partner • Industry Partner • Initial Hub



Why a network?

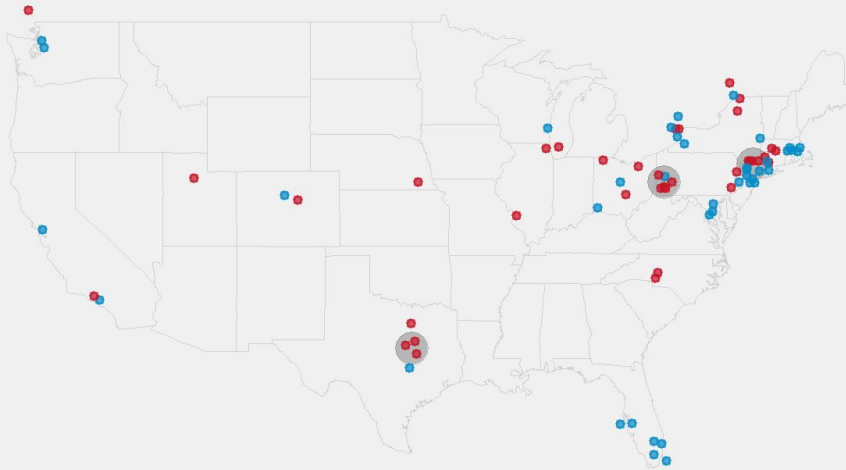
- Decentralized/crowdsourced content creation
 - Ideally sustainable
- Diverse experience and skills
 - Academia, industry and media
 - Key part of review process
- Pre-existing informal network already

Sustainable National Network



Map of Initial Hubs and Confirmed Partners

• Academic Partner • Industry Partner • Initial Hub



Initial Network

- Over 80 initial confirmed partners in academia, industry and media
- 3 initial hubs
 - PIT - Carnegie Mellon, Pitt
 - NY - St Lawrence, West Point, Yale
 - TX - Baylor
- High level advisory board from academia and industry

SCORE with Data



Building a sustainable national network for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

SCORE with Data



Building a **sustainable national network** for developing and **disseminating Sports Content** for Outreach, Research, Education in data science

SCORE with Data



Building a **sustainable national network** for **developing and disseminating**
Sports Content for Outreach, Research, Education in data science

Disseminating Content



ScoreNetwork.org

SCORE Module and Data Repository



Modules can be found at:

<https://modules.scorenetwork.org/>

Data can be found at:

<https://data.scorenetwork.org/>

Module



Educational materials for statistics and data science

Plug and Play into your classroom with flexibility

Open Access (download and edit)

Peer Vetted with DOIs

Pedagogical, Industry, Student (optional)

Following best educational practices

Sports motivated with background materials

Diversity of sports esp. Non-traditional sports

Examples: Published



Outliers using League of Legends data

Correlations with Female Ironman Lake Placid Finisher data

Lacrosse faceoff proportions hypothesis testing

NHL shooting percentages with linear regression

Module Components

Learning Goals

Introduction/Motivational Video

Methods (Statistical/Data Science)

Exercises/Activities

Conclusion

Dataset(s)

Data Glossary

ReadMe file

Module Repository



19 Published Modules

70 more in pre-print

13 different sports among published

Numerous more in pre-print

Searchable by topic

What other modules would you use?

Baseball	>
Esports	>
Football	>
Hockey	>
Lacrosse	>
Marathons	>
Motor Sports	>
Robotics	>
Tennis	>
Triathlons	>

ANOVA (1)
Bernoulli distribution (1)
Binomial distribution (1)
Brier score (2)
Chi-Square Test (1)
Correlation (1)
Difference in two means (1)
Elo ratings (1)
Exponential distribution (1)
Hypothesis testing (1)
Linear regression (2)

Module landing page



[SCORE Module Repository](#) [Home](#) [Modules By Topic](#) [Preprint Access](#) [Data Repository](#) [SCORE Network](#)

[Modules By Topic](#)

[Preprint Access](#)

[Baseball](#)

[Esports](#)

[Football](#)

[Hockey](#)

[Lacrosse](#)

[Marathons](#)

[Motor Sports](#)

[Robotics](#)

[Tennis](#)

[Triathlons](#)

>

>

>

>

>

>

>

>

>

>

SCORE Module Repository

The [SCORE Network](#) Module Repository enables you to search for modules by either sport (along the left), or you can browse [by statistics and data science topic](#). The modules listed in this repository have completed the required SCORE Network pedagogical and industry peer reviews to become a published module.

Interested in submitting your own module? [Click here to find out more information about the SCORE Network module submission process.](#)

You can also access *preprint* modules from various SCORE Network affiliates below (*note that these materials have not yet completed the SCORE Network review process*):

- [Carnegie Mellon University](#)

DOI (Digital Object Identifier)



Permanent and persistent link to modules/data

Makes it easier to discover or search for modules/educational materials

Adds credibility and traceability to professional products

Datasets for class?

Welcome to the UC Irvine Machine Learning Repository

We currently maintain 668 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!

[VIEW DATASETS](#) [CONTRIBUTE A DATASET](#)

Popular Datasets	New Datasets
Iris A small classic dataset from Fisher, 1936. One of the earliest known data sets. Classification 150 Instances 4 Features	Dataset for Assessing Mathematics Learning in Higher Education MathE is a mathematical platform developed under the MathE project ... Classification, Regression 9.55K Instances 8 Features
Dry Bean Images of 13,611 grains of 7 different registered dry beans were taken ... Classification 13.61K Instances 16 Features	Micro Gas Turbine Electrical Energy Prediction This dataset consists of measurements of electrical power corresponding to ... Regression 71.23K Instances 3 Features

TidyTuesday

A weekly data project from the Data Science Learning Community (dslc.io)

variables observations values

kaggle

CMU S&DS Data Repository

The Data Repository curates interesting datasets for use in statistics and data science education. Each dataset is supported by a *story* describing its origin and application, and a set of interesting *questions* that can be answered using the data. This means:

- Every dataset has context in a scientific field, pop culture, or daily life.
- Beyond context, datasets are interesting. They feature more than just a dozen observations from an antiquated scientific study — many feature thousands of observations of dozens of variables, and answer questions interesting to a wide audience.
- Just like in science, some datasets give null results.
- Instructors can easily build lessons and assignments from the suggested questions.

Datasets are organized by broad subject areas on the left, or you can browse a [sortable list of all datasets](#).

Data Is Plural

... is a weekly newsletter (and seasonal podcast) of useful/curious datasets, published by [Jeremy Singer-Vine](#). There have been [381 editions](#), dating from [October 21, 2015](#) to [July 31, 2024](#). To receive future editions, sign up here:

Check out: <https://cmustatistics.github.io/data-repository/>

What about sports datasets?



SportsDataVerse

An open-source sports analytics and data organization.

We provide utilities in Python, R, Node.js, etc.

CRAN Task View: Sports Analytics

Maintainer: Benjamin S. Baumer, Quang Nguyen, Gregory J. Matthews
Contact: ben.baumer at gmail.com
Version: 2023-04-06
URL: <https://CRAN.R-project.org/view=SportsAnalytics>
Source: <https://github.com/cran-task-views/SportsAnalytics/>

SCORE Data Repository

<https://data.scorenetwork.org/>



SCORE Sports Data Repository Home Datasets By Topic Submit a Dataset Data Sources Module Repository

Datasets By Topic

Data Sources

Submit a Dataset

Baseball

Basketball

Combat Sports

Diving

Esports

Football

Golf

Hockey

Lacrosse

Motor Sports

Olympics

Rodeo Sports

>

>

>

>

>

>

>

>

>

>

>

>

SCORE Sports Data Repository

The [SCORE Network](#) Sports Data Repository curates interesting datasets across a variety of sports for use in statistics and data science education. Each dataset has the following properties:

- A *sports question* of interest, with context motivating why the dataset is relevant and interesting to explore.
- A *statistics / data science topic* which the dataset can be used to help teach.
- *Example questions* that instructors can use to help build lessons, handouts, and SCORE modules.

Datasets are organized by sport along the left, but you can also browse [by statistics and data science topic](#).

This repository is heavily inspired by the [CMU S&DS Data Repository](#).

The development of the SCORE with Data network is funded by the National Science Foundation (award 2142705).

SCORE Data Repository

<https://data.scorenetwork.org/>

SCORE Sports Data Repository Home Datasets By Topic Submit a Dataset Data Sources Module Repository



Features over 70
datasets across more
than 30 different
sports

Can search for
datasets by sport
and by statistics &
data science topic

Datasets By Topic

Data Sources

Submit a Dataset

Badminton



Baseball



Basketball



Combat Sports



Cricket



Disc Sports



Diving



Esports



Fencing



Football



Golf



Gymnastics



Handball



Hockey



Lacrosse



Motor Sports



Obstacle Course



Olympics



Rodeo Sports



Running



Skating



Skiing



Soccer



Softball



Datasets By Topic

Jun 26, 2024
Abigail Smith

2018-2023 Badminton World Tour Points Head To Head

HISTOGRAM SUMMARY STATISTICS SIDE-BY-SIDE BOXPLOTS
DIFFERENCE IN MEANS HYPOTHESIS TEST
LINEAR REGRESSION
CONFIDENCE INTERVAL FOR REGRESSION SLOPE

Analyzing wins and points head to head in
singles and doubles in the Badminton World
Tour from 2018-2023.

Hope Donoghue,
Robin Lock

2022 Division III Women's Soccer Results

CORRELATION CHI-SQUARE TEST FOR ASSOCIATION

Division III women's soccer teams game
results from 2022 season

Jun 27, 2023
Jack Fay, A.J.
Dykstra, and Ivan
Ramler

2023 Boston Marathon runners

SUMMARY STATISTICS OUTLIERS Z-SCORES

The data set looks at the 2023 Boston
Marathon results.

Categories

All (56)
ANOVA for means (3)
Bootstrap distribution (1)
Boxplot (1)
Bradley-Terry (1)
Categorical predictors (2)
Chi-square test for association (2)
Comparative plots (1)
Comparing groups (1)
Confidence interval for a mean (4)
Confidence interval for regression
mean (1)
Confidence interval for regression
slope (1)
Confounding variables (1)
Correlation (10)
Data cleaning (2)
Data ethics (1)
Data visualization (4)
Data wrangling (4)
Difference in means confidence
interval (1)
Difference in means hypothesis
test (8)
Distribution description (2)
Elo ratings (1)

What does every dataset include?

- **Descriptive title**, e.g., Women's National Basketball Association Shots
- **Brief description the data background**, potentially include a brief summary of the sports problem and statistical situation
- List of **relevant statistics and data science topics/categories** for tags
- Information about **the source of the data motivating its usage**
- **Description of dataset**: what does one row represent? README file serving as the data dictionary for the columns
- **Example questions** associated with the data

Example dataset page information



Basketball > Women's National Basketball Association Shots

Women's National Basketball Association Shots

CLASSIFICATION

LOGISTIC REGRESSION

GENERALIZED ADDITIVE MODELS

MULTINOMIAL LOGISTIC REGRESSION

NAIVE BAYES CLASSIFIER

DENSITY ESTIMATION

Information about shots during the 2021-2022 WNBA season

AUTHOR
Ron Yurko

PUBLISHED
March 25, 2023

Motivation

The [Women's National Basketball Association \(WNBA\)](#) is the top professional women's basketball league in the world. The league records every shot players take along with contextual information about the shot such as its location, a description of the shot type, as well as the outcome. With this dataset, you can predict the success of each shot attempt to compute the expected value of shot types and compare team decision making.

Questions

1. Build a classification model to predict the shot outcome based on the spatial x,y coordinates of the shot.
2. Create a visualization displaying the joint frequency of shot locations. Do there appear to be any clear modes of frequently taken shots? Create a conditional version of this display by shot outcome. Does the distribution shape vary by shot outcome? (You can also perform a similar analysis by team and shot type).

Data

This dataset contains information about 41,497 shots during the 2021-2022 WNBA season.

The data was collected using the [wehoop package in R](#).

Variable	Description
game_id	Unique integer ID for each WNBA game
game_play_number	Integer indicating the recorded play number for the shot attempt, where 1 indicates the first play of the game
desc	String detailed description of shot attempt
shot_type	String description of the shot type (e.g., dunk, layup, jump shot, etc.)
made_shot	Boolean denoting if the shot was made (TRUE) or not (FALSE)
shot_value	Numeric value of the shot outcome (0 for shots that were not made, and a positive value for made shots)
coordinate_x	Horizontal location in feet of shot attempt where the hoop would be located at 25 feet
coordinate_y	Vertical location in feet of shot attempt with respect to the target hoop (the hoop should be a little in front of 0 but the coordinate system is not exact)
shooting_team	String name of the team taking the shot

Dataset requirements

- The data must be publicly shareable
- If you used code (such as an `R` package) to access the data then include your code with your submission. The code will be publicly available on the [repository's GitHub page](#) for others to view.
- There should be an interesting sports question coupled with a statistics and data science topic to motivate the use of the dataset in educational material.
- The dataset should be in a standard format such as a CSV file and be of reasonable size. GitHub has a file size limit of 100 MB, and large files can be inconvenient for students. We recommend compressing files larger than a few megabytes. Note that gzip compression is a good choice since common tools such as `R` and `Python` feature ways to read `.csv.gz` files directly.
 - e.g., WNBA shot data was compressed and is available to download as `.csv.gz`

This dataset contains information about 41,497 shots during the 2021-2022 WNBA season.

The data was collected using the [wehoop package in R](#).

[wnba-shots-2021.csv.gz](#)