Jessica Hamblin

Spencer Kimball

Braden Wellman

Final Group Project

1. <https://github.com/hamblinjm/crossfit>
2. Image attached as “WorkOutPic.png”
3. Predicting CrossFit Athletes’ Performance
4. Data Set:
   1. Give a list of all the variables in your dataset and an explanation of what they mean

* division: 1= male, 2=female
* stage: 0 = roster (all athletes who signed up), 1 = WoD 1 of that year, etc
* score: expressed in reps or seconds
* scaled: 0 = Rx, scaled = 1. all NA before 2015

I refer to 15.1A as 15.1.1 so that the database field can be numeric  
the scraping process is not 100% successful, so there may be a small number of missing records. Not a problem if you’re summarizing trends. Might be a problem if you’re doing reporting for individuals.

[Athlete profiles](https://drive.google.com/drive/u/0/#folders/0B-JqrzitvsWSZ2NyM1NKd2EwRnc/0B-JqrzitvsWSfmVjY1BmQzVPblZ5WjEzVTktakt2b195ZEpLS1UxZmtDMU5iWERDZ0RFUVU): everything on the athlete profile pages (age, weight, affiliate, team, lift PRs, workout PRs, background, training and diet descriptives).  
athlete profiles can change at any time. 73% of records have a retrieved\_datetime field to make this less ambiguous. The rest were scraped before I thought to do that, but were scraped in March 2015.

* 1. Report the mean and standard deviation of all numeric variables
  2. Report value counts for all categorical variables
  3. Specifically declare which variable is the target (if applicable)

Our target variable is going to be the variable “rank”.

D2: NUMERIC VARIABLES (Count, mean, standard deviation)

|  |  |  |  |
| --- | --- | --- | --- |
|  | count | mean | std |
| FOD\_ID | 30704 | 41319433.5 | 90108133.7 |
| LOCAL\_FIRE\_REPORT\_ID | 9615 | 100.821737 | 700.229485 |
| FIRE\_YEAR | 30704 | 2003.33856 | 6.49607028 |
| DISCOVERY\_DOY | 30704 | 204.798104 | 42.9123905 |
| DISCOVERY\_TIME | 27922 | 1506.69289 | 394.93278 |
| CONT\_DOY | 27507 | 206.067256 | 42.9839394 |
| CONT\_TIME | 27390 | 1576.46444 | 446.121046 |
| FIRE\_SIZE | 30704 | 134.002508 | 2649.77138 |
| LATITUDE | 30704 | 39.2013562 | 1.40222034 |
| LONGITUDE | 30704 | -111.66148 | 1.31776735 |

D3: Categorical variables (Number unique values)

FPA\_ID 30704

SOURCE\_SYSTEM\_TYPE 3

SOURCE\_SYSTEM 6

NWCG\_REPORTING\_AGENCY 6

NWCG\_REPORTING\_UNIT\_ID 43

NWCG\_REPORTING\_UNIT\_NAME 43

SOURCE\_REPORTING\_UNIT 52

SOURCE\_REPORTING\_UNIT\_NAME 62

LOCAL\_INCIDENT\_ID 10763

FIRE\_CODE 10505

FIRE\_NAME 18781

ICS\_209\_PLUS\_INCIDENT\_JOIN\_ID 836

ICS\_209\_PLUS\_COMPLEX\_JOIN\_ID 2

MTBS\_ID 493

MTBS\_FIRE\_NAME 451

COMPLEX\_NAME 49

DISCOVERY\_DATE 4639

NWCG\_CAUSE\_CLASSIFICATION 3

NWCG\_GENERAL\_CAUSE 13

NWCG\_CAUSE\_AGE\_CATEGORY 1

CONT\_DATE 4568

FIRE\_SIZE\_CLASS 7

OWNER\_DESCR 15

STATE 1

COUNTY 117

FIPS\_CODE 29

FIPS\_NAME 29

D4: Target

The target is NWCG\_CAUSE\_CLASSIFICATION (categorical)

Natural 18352

Human 9528

Missing data/not specified/undetermined 2824