Data Source

<https://github.com/BobAdamsEE/SouthParkData/blob/master/All-seasons.csv>

data description and schema

* 10-15 page paper (+ code)
* Canvas submission, one PDF with everything (code can be in a separate file)
* Should cover
  + data source
  + data description and schema
  + data pre-processing required (parsing, filtering, etc.)
  + any bad data issues
  + your Spark algorithm
  + description of any other ecosystem or additional tools
  + output description
  + how did you verify that your output is correct?
  + performance/scale characteristics
  + what would you have done differently if you did this again?
  + conclusions