SI 330 Fall 2019 Project

Goal

The purpose of this project is to have you demonstrate your ability to acquire, manipulate, and clean real-world data both as an individual data analyst as well as in a larger team environment. The intention is that through this project you will demonstrate the skills you have learned in this course, and the rubric for grading is aligned roughly with those skills.

Overview

You're a fresh data analyst out of the BSI and you just landed a gig to work at a skunk works media group which runs a website and a couple of television channels for seeing sports and athletic events. They've employed a team of fresh new BSI's (45 of them, which is a ~ $2.7M investment per year) to help build out content for the company. The company believes there is a market for data driven sports reporting, where expert personalities are replaced with insights from data. The vision articulated by the CEO was that each person would work with different data related to sports/athletics, and the company would build a single portal around that data, then leave dark mode and start to raise funds using this as the seed of the business model. You and the other 44 new hires have one month to build something interesting.

The CEO's Director of Quality Control (me) is going to evaluate your efforts across three dimensions, data gathering, data cleaning, data preparation, and data reporting. At the end of the probationary period (coincidentally this happens to be the end of this course!) you'll be given a grade. The criteria for interesting data by the company is actually quite broad, since part of the appeal is to go beyond the big money maker sports (e.g. NFL, MLB, NBA, and of course, NHL) and hit the "long tail" of people's interest. Badminton? Check. Finnish Dressage? Sounds cool! Little league baseball? Why not!? However, the CEO's vision is that for the sport you choose you must have at least the following attributes:

1. Must contain events which cover a time period
2. Must contain entities (e.g. teams) which have a geography
3. Must have some kind of competition leading to outcomes
4. Must have variation in the above of at least 3 distinct units.

This last item means that there must be at least 3 different time periods (e.g. seasons or events), 3 different geographies (e.g. locations or routes), and 3 different outcomes (e.g. championship games).

Delivery #1 - Pitch

The Director of Quality Control is nervous, this is significant capital outlay for new hires and he's new on the job and wants to make sure everything is looking good. You need to provide him with an 800 word document on the dataset you are planning to look for and how it meets the four criteria above. In addition, the Director recently read a book on building community and decided that everyone is going to have to show up on **October 16th at 8:30am** to share their pitches with one another. Text copies of the pitches should be sent to him via email by midnight the same day with the subject line **OMG I had this great idea!**.

Delivery #2 - Notebook

Information to come at a later date