Zach Lyons

Information Exploration

9/17/18

Problem Set 3

**Monday**

1. This script takes any .png image and turns it into a text image.
   * 
2. Used pip install delorean to install the package.

* Printed out this: Delorean(datetime=datetime.datetime(2018, 9, 10, 16, 48, 25, 978629), timezone='US/Eastern')

1. Instead of using wget, I used curl in a few different ways to get these files. All can be found in the repository.
2. I used the following: grep “Colorado” website.txt

**Wednesday**

Hypothetical Datasets

* Records of BuffOne Card swipes at the rec center. This should include:
  + Student ID
  + # of Hours spent at gym
  + Number/type of intramural programs
  + Number of classes taken at rec center
  + Etc.
* Student Academic Records
  + Student ID
  + GPA
  + Major
  + # of Credit Hours
* Student Registration Data
  + Student ID
  + State of residence
  + Full-time/Part-time status
  + On/Off Campus living status

**Context**

The students at CU are obviously the target audience in this data problem and are also one of the primary stakeholders. Other primary stakeholders include CU regents/management and Recreation Center staff/management. The primary objective here should be to either validate or invalidate the findings they’ve come across regarding gym attendance and GPA. From there on, the rest of this data can be used to analyze current gym use, see how the recreation center system benefits or hurts certain groups of students, and how other variables can impact gym attendance.

**Needs/Vision**

The primary problem to be solved is to see if there is really a connection between gym attendance and GPA. This can be done using a simple correlation between the two types of data. Once we have an answer on that, we can dive into other problems. Assuming that gym attendance really does improve GPA, we can look for other problems to solve related to that. For example, we could look at attendance and GPA stratified by variables like major, on/off campus living, etc. This will give us an idea of groups to target with a marketing campaign that may get them back in the gym.

**Outcomes**

As I mentioned earlier, the likely result of our research would answer the initial research question regarding a gym hours/GPA correlation. However, if it all goes well, the next step would likely go back into a marketing campaign targeted specifically at groups in the CU community who do not regularly come out to the recreation center. Once we send this out, we can continue to dig into the BuffOne Card swipe data to see if it made any sort of impact in attendance within these communities.

**Friday**

**Moneyball**

*Moneyball* is a book by Michael Lewis published about the 2003 Oakland Athletics, a team who was hit harshly by the realities of being a small-market baseball club. All their heavy hitters from the 2002 season left to join large-payroll clubs like the Yankees. This left the A’s in a precarious spot, as they had lost all of their best players and had limited financial resources to replace them. General Manager Billy Beane turned to analytics, where his front office staff attempted to use databases of player statistics to find players that the market “undervalued”. A key stat they homed in on was On Base Percentage. Their studies led them to believe that this was the stat most correlated to high run scoring, and as a result their new team of “undervalued” players made it back to the playoffs.

**Astroball**

*Astroball* is a book like *Moneyball*, but the team focused on in this iteration is the Houston Astros. Their approach was a little different than the one used by Billy Beane in Oakland. As opposed to the A’s, the Astros were one of the worst teams in baseball before they started their analytic revolution. In 2011, after another 100-loss season, the Astros turned to Jeff Luhnow, the former analytic department manager for the Cardinals. They took the same approach as the A’s, but applied it top to bottom in their organization, as opposed to just in the free agent market. They built out massive databases of their own analysis that extended to Minor League players, not just Major Leaguers. The purpose of this was to train a machine learning system to project the future value of Minor League players. They then turned all their roster-building strategies into future-minded approaches. If the current value of a player they were receiving did not exceed that of the future value they were giving up in Minor League players, they didn’t make the trade. The plan worked, as the Astros won the 2017 World Series.

**Netflix**

My last example is a fairly ubiquitous one, but still important to mention. In 2006, Neflix was looking for ways to improve its movie rental through mail business. The company was sitting on troves of data from users, over 100 million ratings. Their goal was to improve the recommendation algorithm by a mere 10 percent. To do this, they released this data to the public, and let data scientists work with it to improve recommendations. While no team could crack that 10% threshold, the company handed out a few million-dollar prizes to teams that came up with worthy improvements. I can only assume the research from this study was one iteration of the current system, which relies on a thumbs up/down approach.