Name: Jacob Paul

Collaborators: Taylor Lawrence

Problem Set 3

**Monday**

1. In this case, the script prints out the GitHub octocat mascot. The script has has the ability to convert any image into ASCII art by simply passing the image as an argument. Screenshot in repo (asciify\_screenshot.png)

2. Command to install delorean: "pip install delorean"

After running the commands from the instructions, this was the output:

"Delorean(datetime=datetime.datetime(2018, 9, 10, 16, 19, 29, 224914), timezone='US/Eastern')"

3.Command used: “wget http://www.colorado.edu/”

index.html is the output of the command and is included in repo.

4. Command used: “grep Colorado index.html -c”

There were 17 lines that contained "Colorado"

5. Command used: “grep Colorado index.html > search\_results.html”

**Wednesday**

Dataset 1: Dataset with all BuffOne card swipes for entry into the gym with all student IDs with timestamps.

Dataset 2: Data on all students, their year, GPA, Major.

Dataset 3: Anonymous survey that asks students about their gym usage, specifically how often they go to the gym, when they go, what they do while they are there, if they go with other people, why they go and how long they are there.

Context:

* Stakeholders: CU Leadership, Facilities Management, Students, Marketing Department
* The goal is to discover more information about student gym habits to optimize facilities and facilitate higher average GPA’s
* The target audience is the CU leadership team who will make the decision on what changes might get made and if they get implemented

Needs:

* CU would be able to fix an un-optimized gym system and make it work to the benefit of students.
* The results will point to factors of students’ gym attendance habits that result in a higher GPA. By knowing these factors, CU will be able to change aspects of the gym experience to be more conducive to the activities that lead to higher GPA’s
* We will have knowledge of CU students’ gym habits and how that relates to GPA

Vision:

* Some techniques that could be used are statistical modeling (correlation, regression, etc.), thematic analysis on the survey data
* When finished, we will have an idea of how well GPA relates to a student’s gym attendance habits and what types of activities students go that may correlate to a higher GPA

Outcome:

* The results will be used to validate previous findings that pointed to higher GPA’s for students who attend the gym regularly. They will also be used to find out what CU students do most at the gym and which of these activities CU should focus on when implementing changes to the gym facilities.
* When we are done, leadership will presumably make a series of decisions of whether or not to go ahead with the changes that may have been prompted by our data
* The Office of Analytics or Facilities Management will most likely be responsible for what happens to the data when we are done.

**Friday:**

Example 1: Fraud detection for credit cards

1. The problem: credit cards were getting compromised after getting stolen when being used at a compromised store or website. In the early 90s, credit card companies were using physical elements on cards to help prevent fraud, but it wasn’t working. Criminals were still able to defeat these measures.
2. Transaction data from credit cards became available and easier to access as payments became more electronic.
3. Initially, companies like HNC software designed software like Falcon Fraud Manager to use early forms of data analysis to create profiles of cardholders and compare transactions to those profiles. Later on, neural network models and others came into use for fraud detection.
4. Since the advent of this technology, credit card payments have become much more secure and fraud is much less of an issue than it was 20 years ago.
5. Today, we have much more computing power and more advanced models available to us. Many companies utilize supervised and unsupervised machine learning techniques to group transactions in respect to whether or not they’re fraudulent. Transaction location and amount are commonly used variables to train models like these. If I were to tackle this problem today, I would explore what other variables may be able to predict fraud more accurately such as time of day, daily/weekly/monthly spending patterns, types of goods/services being purchased, and usage of other devices/services at the time of the transaction.

Information source: <https://www.kdnuggets.com/2014/03/evolution-fraud-analytics-inside-story.html>

Example 2: UPS Navigation System

1. UPS needed to find the fastest and most efficient way to deliver packages to customers without relying on driver knowledge of a given area.
2. UPS had GPS tracking data of routes as well as data as to when the package was shipped and if they met delivery commitments.
3. By doing using the data they had gathered, UPS created the ORION software, which is a navigation system that takes into account this historical data UPS had gathered as well as map data to create the most efficient routes for delivering packages.
4. UPS was able to increase the efficiency of their delivery system while also saving significant amounts of fuel and therefore saving money.
5. UPS came out with this software only 5 years ago, but if I were to tackle the problem today, I would do so by diving deeper into using machine learning models to determine what routes may have traffic on them and consistently make deliveries later than expected.

Information from: <https://www.forbes.com/sites/alexkonrad/2013/11/01/meet-orion-software-that-will-save-ups-millions-by-improving-drivers-routes/#7135aae14fc7>

Example 3: San Francisco Bike injuries

1. The amount of people that were dying and getting seriously injured as a result of an accident while on their bike was on the rise in SF.
2. In order to generate data to influence policymaking on this issue, the city developed TransBase to map and visualize traffic incidents in the city.
3. The team used this data to develop what they called the “high injury network” which outlined the top places where injuries happen. They used mapping tools like ArcGIS to map these areas and publish this data.
4. Using simple analysis tools, they were able to determine that most injuries occur at a small fraction of intersections. This allowed the city to focus their resources on a small number of intersections rather than blanketing the city with costly measures.
5. If I were to tackle this problem today, I would add a survey or interview component. It would be interesting to merge the injury data that the city collected with real experiences of those who commute by bike and/or have been involved in an accident while they were biking.

Information from: <https://medium.com/@abhinemani/data-driven-policy-san-francisco-just-showed-us-how-it-should-work-c7725e0e2b40>