Assignment 3

Monday:

1. This script runs the python file, which contains an image of octocat.



1. I used the command python -m pip install delorean.

When I ran the code, the command line printed: Delorean(datetime=datetime.datetime(2018, 9, 14, 0, 52, 50, 472667), timezone = ‘US/Eastern’)

1. After painstakingly installing wget and making it one of the Windows command line environment variables, I used the command wget to download <http://www.colorado.edu/>.
2. To count the number of lines with the word Colorado, I used the command find “Colorado” index.html | find v/ c/ “” which told me there’s 18 lines.
3. To save the search results into search\_results.html, I used the command findstr “Colorado” index.html > search\_results.html.

Wednesday

Some hypothetical data sets that I would use to solve this problem would be:

1) a dataset that contains the records of which students attend the gym and the frequency with which they attend, ideally one that contains their CU student IDs

2) another data set that has students’ CU IDs and their academic performance information (like GPA, grades, etc)

3) another data set that might contain what kind of things students engage in while they’re at the gym (like weight lifting, swimming, basketball, etc), ideally also containing the names or CU IDs of the students whose responses are being recorded or a survey of the student body (including students who attend and don’t attend) to determine their attitudes towards the CU gym.

The CoNVO method to solve this problem would go something like this:

Context: They’re trying to figure out the current state of gym use and its potential effects on students to eventually determine how they might target specific facilities or programs to improve attendance. This goal is describing a desire to increase gym attendance in an attempt to increase students’ academic performance. I imagine the stakeholders would include those people who are initiating this project, the people who work at the gym (if the potential solutions resulting from this project end up affecting the way they’re required to do their jobs), the people in charge of running the gym, the people who have access and the ability to share the data sets we need to answer this problem (which will probably be proprietary and it’s unlikely we’ll be given free access to it), and those students are already regularly attending the gym as well as those who we’re trying to get to attend.

Some questions I would ask my stakeholders before starting this project would be: what particular questions about the current state of gym use are you interested in figuring out? How do you define “the current state of the gym”? What about its current state are we interested in? What sort of potential effects of the gym are we interested in (besides student GPA)?

Needs: The specific questions we’re trying to address are: can we observe a trend between regular gym attendance and student GPA in CU Boulder students who attend regularly? If so, what sort of trends can we observe about these students from their data? What sort of activities make these students want to attend regularly? What are the most popular activities at the gym? What do students like/not like about the gym? For those students who don’t attend, why don’t they attend?

I would use these datasets to look for trends that would answer the questions listed above and leverage the data to figure propose tactics that the gym could use to increase attendance.

Vision: My analysis tool of choice would probably be Tableau and Tableau Prep because I’m most comfortable with these tools at the moment. I suppose that Python can be used as well if needed to run more sophisticated regressions. I would likely use Tableau Prep to clean the data and join data sets. I would then use Tableau visualizations and tables to compare data using different dimensions. Once I find concrete solutions to the questions, I would create dashboards or a data story to present those findings to my stakeholders and offer some solutions that are backed up by the data. I think that using a visual presentation of the data to back up the findings also has the added bonus of improving stakeholder comprehension and credibility of the findings.

Outputs: Ideally, what we have when this is all done is a Tableau dashboard and series of suggestions that the CU gym can use to get more students in the door as well as improve their existing facilities to retain those students who already attend. The stakeholders who commissioned the project can then handover the insights to the gym managers and marketing teams so they can go to work figuring out how they want to act on the findings.

Friday

1. What was the problem?

a) In 1663, John Graunt recorded information about mortality in an attempt to design an early warning system for the bubonic plague that was ravaging Europe at the time.

b) In 1880, the US Census Bureau had a problem because it estimated that it will take it 8 years to crunch all the data collected in the 1880 census, and it is predicted that the data generated by the 1890 census will take over 10 years, meaning it will not even be ready to look at until it is outdated by the 1900 census.

c) The Ministry of Posts and Telecommunications in Japan starts conducting the Information Flow Census, tracking the volume of information circulating in Japan to determine its impact on the public.

2. What data was available?

a) John Gaunt compiled statistics about the citizens of London over a 70-year period. This included the Lists of Mortality (lists that were published in newspapers of how many people died during the years that the plague was ravaging England), along with christening records from churches and data from an area of rural England.

b) 1880 US census data.

c) The census tracks the “amount of words” produced by the media and uses it as the unifying unit of measurement across all media.

3. What methods were used?

a) Gaunt had grouped together the facts that were displayed in the 70 years of records and noted the comparisons of findings for different populations. His methods were not all together that clear (he seldom included calculations and sometimes omitted important information).

b) In 1881, Herman Hollerith creates the Hollerith Tabulating Machine, which uses punch cards to reduce the 10 years worth of work to 3 months.

c) Tracking the amount of words produced by the media.

4. What did they find?

a) He offered the first well-reasoned estimate for London’s population and he used evidence from medical records to refute the idea that plague spreads by contagion and that it occurs early during the reign of a new king.

b) They discovered a better way to track the data.

c) The 1975 census already finds that information supply is increasing much faster than information consumption and in 1978 it reports that “the demand for information provided by mass media, which are one-way communication, has become stagnant, and the demand for information provided by personal telecommunications media, which are characterized by two-way communications, has drastically increased…. Our society is moving toward a new stage… in which more priority is placed on segmented, more detailed information to meet individual needs, instead of conventional mass-reproduced conformed information.”

5. If you were to tackle the same problem today, what would you do differently and why?

a) I would try to obtain records that might indicate the geographic route by which the plague spread to try to establish a common link between the cases.

b) In 1881, I would have probably done the same thing. In modern day, AI and other modern tools could organize this data in a matter of hours.

c) I would probably compare the amount of information produced in the media with the amount of information that the human brain is said to be able to process. This would be a very long and difficult research project, but it would essentially determine the same thing that the Japanese Ministry of Posts and Telecommunications already found.