Data Science Project Full Proposal

Group Members:

Erik Ronning - eronning
Bruce Nguyen - blnguyen
Ayan Tazhibayev - atazhiba

Capstone:

No group member is taking this class for a capstone.

Data:

- **Menu data**: all of the menu information for the Ratty in the past year
    - The menu data includes all items being served at the Ratty. The data here has yet to be cleaned but we plan to just take into account that tracking specific menu items and associating them with the traffic could be complicated as the Ratty serves a multitude of entrees. We may opt to clean the data by the type of cuisine that is being served as opposed to the specific menu item. Additionally, we have to figure out what items are constantly recurring on a daily basis and ignore them in our analysis.
- **WiFi data**: all of the WiFi information for each of the access points in the Ratty for the past year
    - The wifi data here just shows us how many people disconnect and connect to the wifi at a specific access point in the Ratty. There are timestamps for these connections and disconnections and additionally mac addresses. First off, we have to combine all the access points into one file to get the total count of WiFi connections in the Ratty. Additionally, cleaning the data here will be tricky because it depends on how we wish to conduct our analysis. Since we are going to use this data as a proxy for how many people are in the ratty at a given time on a given day, we need to heavily clean the data, given that it consists of just disconnection and connection timestamps. We are planning to loop through the data and calculate how many people are currently in the Ratty in a given time interval. An example would be how many WiFi connections are in the Ratty at 12pm on Monday. How we could calculate this is add the connections and subtract the disconnections from 11am-12pm from the 11am count. This implies that the 7:30am start time will have 0 connections, hence we plan to start analysis at 8:00am to reconcile this issue.
- **Weather data**: weather information for Providence for one year back
    - This includes temperature, wind speed, rain, and snow. This data has technically been cleaned already as it was taken from another API. The variables included here are what we think would be relevant in explaining Ratty traffic. The only other cleaning that is required is to remove dates that are not necessary for our project, such as times when the Ratty is closed.
- **Course times data**: class information and allocations (class sizes and times)

    - This data doesn't really need to be cleaned other than removing class times that fall outside the timeframe of the ratty. We have used it to create visualizations to help us understand what the flow of traffic will look like during a typical day. At this moment in time, we've only used it to formulate hypotheses about what our analysis will reveal. If it is going to be used in our statistical model, it may need to be reformatted but that shouldn't be too difficult.

Methodology:

The same data from the beginning of this project will continue to be used. Some of the data has just been received after spending time working with Brown CIS to get access. So far the data has just been parsed in using Python and JavaScript scripts which have begun to start processing the data (placing information in an easy to access format). So far, the course data for Brown along with the WiFi data in the Sharpe Refectory have been looked at. The question we explored being how the class schedule of students affects the current traffic in the Ratty. Visualizations have been created in order to display this information and clearly portray a suspected correlation between the WiFi data and the course information.

The next step of investigation entails incorporating the weather information that has been gathered, along with the WiFi data, in order to attempt to see if there are any suspected correlations between Ratty traffic flow and the weather. This information could be looked at in several ways. One path would be to explore weather and traffic flow. If there are fluctuations in WiFi connections to the Ratty during the timeframe of specific weather patterns, it may be worthwhile exploring the weather data and Wifi data more in depth. For example, the number of connections could be looked at while the weather is poor(snowing or raining). Another method of looking at the data could be seeing changes in WiFi traffic following or prior to the weather data being viewed. An example would be looking at the number of connections in the Ratty after the weather has gone from bad to good. Investigating these correlations could be done by doing statistical tests or even by attempting to visualize the information (perhaps overlapping weather with WiFi connections on a graph in some way).

Task List:

- **Clean out all invalid dates**, for all datasets, such as holidays or summer as well as clean out invalid times such as the times of day when the Ratty is not open (2 hrs -- by 4/20)
- **Parsing all of the WiFi data** from each access point (30 min -- by 4/20)
    - Aggregate all of the WiFi access point data into one parseable file (30 min)
    - Generate a script capable of parsing all WiFi data and bucket the WiFi connections into specific times of the day, perhaps 15min windows? Details are above in Wifi Data. After bucketing, average our values by given weekdays so we can get data on what we could expect to see on average on a given weekday at a given time rather than on a given date at a given time. (4 hrs -- by 4/20)
    - Take a look at weather in relation to bucketed connections on given dates (5 hrs -- by 5/5)

- **Parse all of the Menu data** gathered from the API (2 hrs -- by 4/20)
  - Bucket similar menu items together where the key is the menu and the value is a list of dates (2 hrs -- by 4/20)
  - Take a look out menu items in relation to bucketed connections on given dates (5 hrs -- by 5/5)
- **Parse all of the Weather data** from the CSV data set (30 min -- by 4/20)
  - Bucket the weather information based on time frame for given dates (3 hrs -- by 4/20)
  - In each bucket, have a basic boolean weather condition variable that indicates whether or not weather is bad or good. (1 hr -- by 4/20)
- **Parse all of the Course Times** data (2 hrs -- by 4/20)
  - Bucket information based on class times. Same class times go in the same bucket, making the assumption that people go to class. (1.5 hrs -- by 4/20)
  - Take a look a start/end points of class times in relation to number of people in the Ratty, perhaps creating a weight? (5 hrs -- by 5/5)
- **Generate more visualizations** from all of the parsed and cleaned data
  - In order to understand flow of traffic create new visualization and recreate old visualizations.
- Being able to **create a linear regression mode**l for predicting how many students will be in the ratty dependent upon certain variables (5 hrs -- by 4/20)
  - So far we have weather information, menu information, class start times, class end times, wifi connections at the ratty
    - With this data, we need to be able to figure out what variables would be feasible and reasonable to include within our model
    - Some variables that would make sense at the very moment are, time of the day, what day it is, good or bad weather, peak meal times: breakfast, lunch, and dinner, expected amount of free students(given by class times).
  - Clean our datasets accordingly so that we can appropriately incorporate them into our linear model.
- **Implement a basic machine learning algorithm** that will roughly predict the flow of traffic contingent upon our data (5 hrs -- by 5/5)
  - This is one of the lower priorities on our list, but if we have time, it would be valuable to implement a more complicated model other than a linear regression. Possible choices: Baeysian, Clustering, Decision Trees, etc.

Deliverables:

- 75%: **Understand flow of traffic** in the Ratty (capable of visualizing flow)
  - Effectively visualizing specific variables that we think would affect the flow of traffic in the ratty.
  - Using these visualizations to identify trends in the flow of traffic.
  - Hypothesizing (but not statistically proving) what variables are relevant in explaining the flow of traffic.

- 100%: Able to **estimate the flow of traffic** in the Ratty (requires a statistical model)
    - The goal here would be able to train a statistical model on our dataset with specific relevant variables
    - This doesn't imply that we want to create a statistically significant and suggestive model but rather we want to be able to train and create a model given our set of data
    - For contextualization, an example model that would suffice would be a model that could how many students are in the ratty (perhaps in sample) within a specific and lenient margin of error.
- 125%: **Capable of determining the significance of flow of traffic of a given menu item** based on number of people in the Ratty throughout the day and provide recommendations to BuDS (more rigorous analysis of our data)
    - Identify a model that is statistically significant in explaining in sample data perhaps maybe even a machine learning algorithm
        - If possible find a way to be able to create a model that can predict out of sample data..perhaps using past years as test data
    - Using our model, suggest menu items that would increase the user flow in the Ratty.
    - Suggest to the ratty how much food to prepare given a characteristics such as the weather
    - Any sort of real world applications that can be suggested, contingent upon our analysis

Backup Plan:

It doesn't seem to be an issue that we won't be able to generate a model for predictions. What is concerning to us is being incapable of generating any significant models which provide a reasonable and loosely accurate estimation(our 125% goal). In the case that this happens, we will do two things: Emphasize the process which we went about to create our model and two, generate a lot of visualizations that do our best to demonstrate the flow of traffic in the Ratty.

The idea here is showing that we are able to analyze a large dataset and understand what it can and can't tell us with our given model. An important part of data science is interpreting data correctly, so another thing we could do in this case is criticise our model. We could talk about possibilities of why our model fails to be significant. This could be the data or the way we constructed our model or even both. This will require a lot of speculation and hypothesizing on our end as to why are weren't able to come out with concrete conclusions. Another great thing we could do is suggest what we would do in the case that we were given more time to complete this project. This could be anything from exploring other relevant datasets or going with a different statistical model.

In addition, we will also generate other relevant visualizations to understand trends in the ratty traffic. Some examples could be visualizations that display the most popular menu items or the effect of weather on flow of traffic. We will also generate a detailed write up regarding the trends that we observed through this dataset that are relevant and/or interesting. The write up would also detail the tasks that would be left to finish the project goals as well as what parts of the project were completed. In essence, completing our 100% goal seems very feasible while the 125% is a bit of a stretch. In the likely case that we don't complete our 125%, there is still a lot we could do to create a substantial final project.