Some overall comments:

* need to give more detail about the strata, and the samples of clusters you constructed in each stratum -- more like the discussion we had in class when you presented.

* see notes below about footnotes (I don't prefer) vs list of references and citations from the paper (I prefer)\

* what will go in the appendices? -- extra calculations, summary statistics on each item in your observational protocol, and anything else that should somewhere in the report but may not be part of teh "main story"

Accuracy of PAT Bus Time Tables

At the Forbes and Morewood Intersection

Section 1: Introduction

Many students and Professors at Carnegie Mellon rely on the public transit system to get to work, however students frequently complain about the PAT bus system. The most common complaints are late buses, inaccurate schedules, and the frustration that occurs after waiting for a bus only to have multiple buses of the same route arrive at the same time[1]. Waiting wastes time, causes frustration, and in the long run could lead commuters to choose to find a way to travel that doesn't involve public transportation. To aim of this study is to first measure the degree to which these complaints are accurate, and if buses are systematically late develop a model for predicting expected arrival time.

This study will be build on a strong general literature base on public transportation and investigate the accuracy of bus time tables for the Forbes and Morewood intersection which is the most commonly used bus stop for commuters at Carnegie Mellon University. Bus departure times will be observed and compared to posted bus schedules. A number of potential factors that influence bus punctuality will also be measured including the weather, the time of day, and the level of light. Using these factors and the information collected on bus arrival times, a model will be created to predict when a bus will arrive given the scheduled arrival time.

Section 2: Methodology

2.1 Target Population and Frame

The target population is all of the buses which stop on campus that CMU students use to get to school. Our sampling frame is the list of port authority buses that stop at the Forbes and Morewood intersection which is the most commonly used bus stop by CMU students. The sample population is 61A, 61B, 61C, 61D, 67 and 69 that stop at the Forbes and Morewood bus stops.

These two stops are adequate for our purposes for several reasons. First, the other Forbes stops (such as Hamburg and Beeler) are very close to the Forbes and Morewood stops. If the bus is late at Hamburg, then it will likely be just as late 100 feet down the road at Forbes and Morewood.

Second, the other stops close to campus (such as Fifth and Morewood, and the Wilkins stops) are used by a substantially smaller percentage of the student population and mostly for the purposes of shuttling to non residential locations. The areas in Shadyside in which students live are well serviced by the faster and more reliable loop buses.

Buses moving up and down fifth will also be independent from buses moving up and down Forbes because many travel through different neighborhoods and thus face different traffic

[1] Mishalani, Rabi G., Mark McCord, and John Wirtz. "Passenger Wait Time Perceptions at Bus Stops: Empirical Results and Impact on Evaluating Real-time Bus Arrival Information" *Journal of Public Transportation, Vol.9, No.2, 2006*

I prefer you to break these out into a separate "references" section and refer to them in the text. Examples of text referencecs AND format for reference section at http://www.library.cornell.edu/resrch/citmanage/apa ... please follow these examples

patterns. This would lead to trouble in data analysis because the independent routes could lead to a bimodal distribution, hurt the accuracy of any inference we would like to make about how late buses usually are, or increase the number of man hours beyond what is feasible for our group.

Due to the nature of public transit, the decision was made to use cluster sampling because attempting to observe specific buses would be nightmarishly difficult. Instead of sampling individual buses, the bus schedule was broken into one hour increments of potential observation. These hours of observation were then stratified into groups of similar hours: the morning rush hours, lunch hours, afternoon, and the evening rush hours. The decision was made to exclude extremely early buses and night buses because they are outside the scope of normal commuting times and measuring them would either require more manpower than the team had available or reducing the number of observation hours for the more interesting strata.

please give full details -- which hours were the morning hours, the lunch hours, etc...

A sample was drawn by taking all of the potential hours of observation within a stratum over the two week period and assigning them a number in chronological order. A random number generator[2] was then used to generate a set of numbers to determine which hours would be included in our sample.

2.2 Sample Size

Given that the sample is stratified, clustered and sampled without replacement the sample size calculations are as follows:
We assumed that the number of buses is going to be similar across all strata and an arrival rate of 12 buses per hour
Previous research indicated that a standard deviation of 5 minutes is normal[3]

$$N = 2\ (Weeks) * 7\ (Days\ per\ week) * 12\ (Hours\ per\ day) * 12\ (Buses\ per\ hour) = 2016\ buses$$

$$n_0 = \frac{z_{a/2}^2 * (SD)^2}{(ME)^2} = \frac{1.96 * 5^2}{.5^2} = 384.16$$

$$Sample = \frac{N * n_0}{N + n_0} = \frac{2016 * 384.16}{2016 + 384.16} = 322.6729$$

This number is the number of buses, not hours, so we divide by 12 yielding ~27 hours of observation.
We then inflated by 20% to account for clustering effects: ~32 hours of observation
Since we have four strata, 8 hours will be sampled from each stratum.

[2] http://www.random.org/integer-sets/ neat!
[3] Macaffe, K. Bus Punctuality Statistics GB: 2007. Great Britain: Department for Transport. http://www.dft.gov.uk/pgr/statistics/datatablespublications/public/buspunctuality/buspunctuality07.

Section 2.3 Observation Protocol

Weather condition: Sunny, Rainy, Snowing etc.

Light/dark level: Day, Night, Dawn/Dusk

Road condition: Dry, wet, icy

Inbound/Outbound: Is the bus at the inbound or outbound bus stop at Forbes/Morewood

Date, day of the week & hour of observation

Bus number/route

Scheduled Bus Departure Time:

Observed Bus Departure Time:

Difference between observed and scheduled departure time: Value of the difference between the scheduled and actual departure time, positive for a late bus, negative for an early bus.

On Time: Yes/No. Yes if the bus departs no more than one minute early or five minutes later than scheduled[4].

why are these footnoted?
Please discussed what you
learned from the references.

Show/No show: defined as being so late that arrival time is within five minutes of the next scheduled bus of that type[5].

Highly unusual conditions - these will be written down and are meant to include notable and unusual circumstances like a broken water main, major traffic accident, etc. This may be used to identify outliers.

Special events and other planned conditions.

Rush hour: Yes/No

Clustering: Number of other buses of the same number that arrive within 2 minutes of each other

2.3 Data analysis

-Yet to come

ok

[4] Macaffe, K. Bus Punctuality Statistics GB: 2007. Great Britain: Department for Transport. http://www.dft.gov.uk/pgr/statistics/datatablespublications/public/buspunctuality/buspunctuality07.
[5] Ibid

Section 3: Results
-Yet to come

ok. It would be helpful to know what analyses you may try.

Section 4: Conclusion
-Yet to come