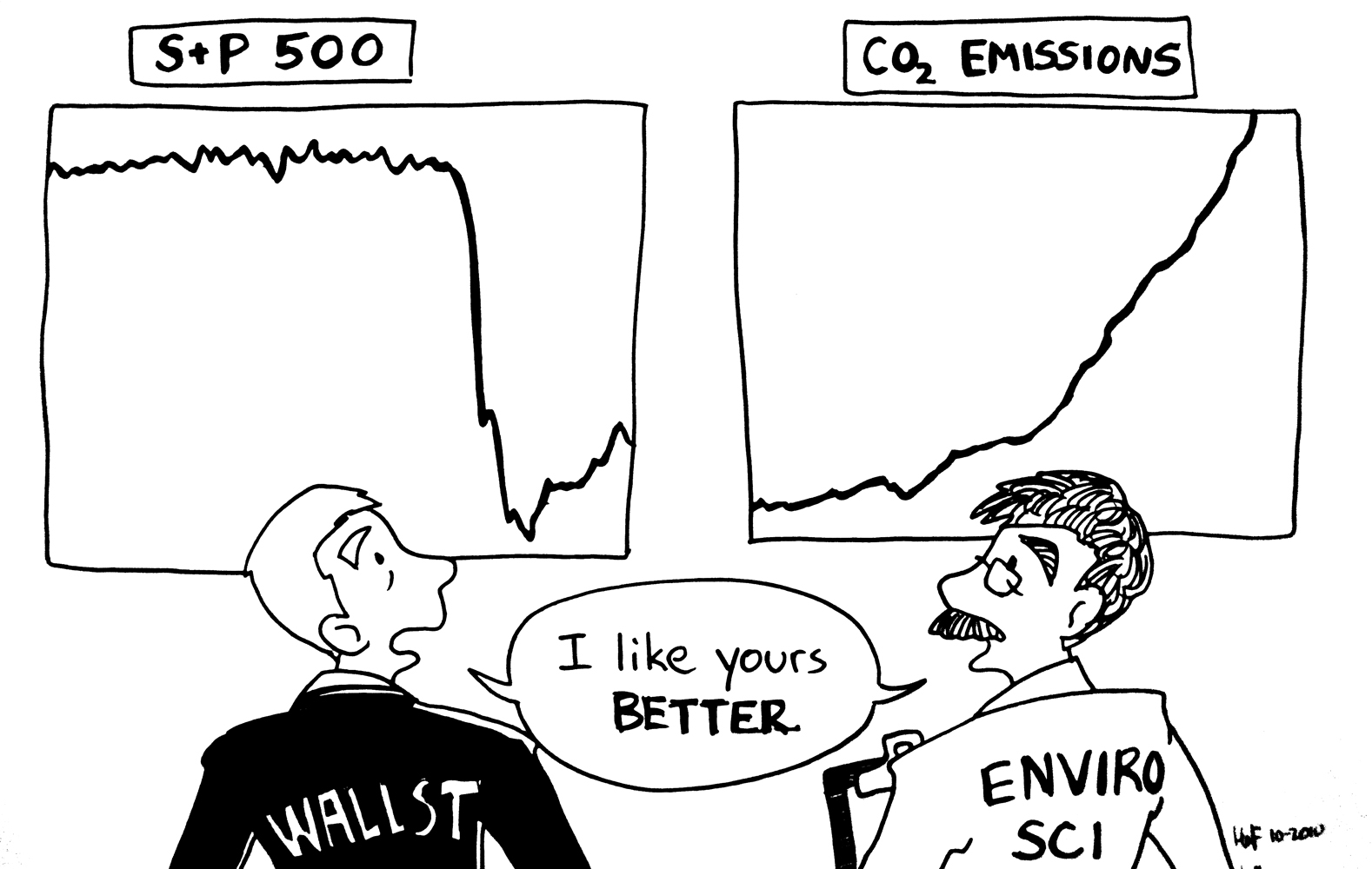
**Carnegie Mellon University, Heinz College**

**Business Intelligence & Data Mining with SAS Suite**

**Fall 2012 (94832 Mini 1)**

**Group Project**

**(Due Oct/11/2012)**

****

**Guidelines for working effectively in a team**

Business activities involve group effort. Consequently, learning how to work effectively in a team is a critical part of your business education. The documents we distributed in class contain a set of common rules to consider adopting for your teams.

**Peer Evaluation**

After each project, we will ask for feedback. Each person must evaluate all team members, considering their individual contributions to the total team effort. This peer evaluation form will be used as an individual evaluation and will **not** be shown to your team members. The form will give us additional information about people's participation that occurred outside the classroom.

**Project Overview**

Now that you have some knowledge about BI and DM, it is time to use the skills for some real work!

We provide three different levels (*Conceptual*, *Hands-on*, *Challenging*) of tasks for you to choose.

**1. Choice 1 (Conceptual):New Trends in BI and DM**

* Identify at least 5 new trends in the BI/DM area. For example, such trends may include (but are not limited to) any new BI/DM applications in Health IT, Social Media, Marketing, Finance, Accounting, Non-profit organizations, High tech industry, Entrepreneurship, Education, Arts and Entertainment, Government, …
* In this task, you need to explain in details: What exactly is the trend? What is its relationship to BI/DM? Is it useful/useless? Why? What is the motivation behind it? What is the advantage and disadvantage? How would you improve in the future?

**2. Choice 2 (Hands-on): SAS EM Data Analytics**

* Find any interesting data source provided in SAS EM (SAS Table, Metadata Repository --> Shared Data --> Libraries);
* Explore the data set, find something interesting (eg, creating interesting plots of the data);
* Use your exploration to help form a hypothesis/question you’d like to answer (Why it is interesting for business/policy/society?...);
* Answer that question by using the BI/DM methodologies we learned in class.

**3.** **Choice 3 (Challenging): Creative Data Analytics**

* Find any interesting data set (I will provide you a list of data, eg, Stock market data, NBA statistics…);
* Explore the data set using any tools (SAS, R, STATA, WEKA, EXCEL, SQL SERVER, MATLAB, C/C++, JAVA, …), find something interesting. You can implement your own data visualization;
* Use your exploration to help form a hypothesis/question you’d like to answer (Why it is interesting for business/policy/society?...);
* Answer that question by using the methodologies learned in/outside class. You can code your own model. You may also want to join together different datasets to discover interesting patterns that may emerge.

**Note:** This is a challenging and also open-ended project, with no pre-defined “correct” answer. It is up to you to locate a data set (or any new trend) that is interesting and possible to analyze in a meaningful manner. Projects that combine multiple data sets and multiple BI/DM methodologies will receive higher grade.

**Data Sources**

* [Infochimps](http://www.infochimps.com/) (many, many data sets available)
* [NYC Open Data](https://nycopendata.socrata.com/)
* [ScraperWiki](https://scraperwiki.com/) (has data from the web that people have “scraped”)
* [Music Data](http://webscope.sandbox.yahoo.com/) Includes user ratings of artists on Yahoo Music
* [Movie Data](http://www.grouplens.org/node/73) 100,000 ratings from 1000 users on 1700 movies.
* [Google Insights](http://www.google.com/insights/search/)
* [Google N-Grams](http://books.google.com/ngrams)
* [World Bank](http://data.worldbank.org/)
* [Yahoo Finance](http://finance.yahoo.com/) (download csv files of stock prices)
* [NFL Play by Play Data](http://armchairanalysis.com/nfl-play-by-play-data.php)
* [Baseball](http://www.baseball-reference.com/) (offers CSV files of many stats)
* [NBA](http://www.basketball-reference.com/) (offers CSV files of many stats)
* [Research Quality Data sets from Hilary Mason](https://bitly.com/bundles/hmason/1)

**Deliverables**

Your report must contain the following sections:

1. Description of the data

* tables
* variables available
* any variables you construct by combining or transforming variables

1. Exploratory plots and tables

* describe the data visually
* show relationships between key variables
* anything interesting that you found

1. Your Research Question and Your BI/DM Methodology

* how you decided on it
* how you will answer it
* what the data say about it

**Note:**If you are working on a conceptual task, then your report need to contain at least five sections, each providing details on one BI/DM trend.

**Note:** Cite the exact data sets that you used for your study (and link to them). This allows others to verify your findings.

**Grading**

* The grade will be determined by three components: (1) peer evaluations (within and between group); (2) your class presentation; (3) your final report.
* Higher grades will be given to higher level choices.
* Higher grades will be given to analyses that analyze bigger data sets compared to small ones.
* Higher grades will be given to analyses that use two or more datasets. e.g. you could combine NYC data at the ZIP code level, or US data on the State or County level.
* Higher grades will be given to reports that indicate a non-trivial level of analysis.
* Higher grades will be given to individuals who provide critical insights to other teams during the Q&A process.

**Group Presentation Guideline: (15 mins presentation + 5 mins Q&A)**

* *Slides* – The quality of the slides. Do they get the points across? Are they appealing and attention getting? Do they facilitate an effective presentation?
* *Presentation* – The quality of the delivered presentation. How were the points delivered? Was the communication persuasive and informative? Did the tone grab and keep the audience’s attention? Did the content flow well? Were the speakers effective in delivering the content?
* *Content* – Was the content accurate, detailed and relevant? Did the content reflect an understanding of the topic and a high degree of analysis? Did the content reflect appropriate frameworks and analytical lenses?
* *Questions* – How did the group handle questions? Were they prepared with good answers? Did they handle the questioners with respect and with an eye to convincing the audience of their views (but while acknowledging good points made by audience members)? (*It is recommended for the audience to ask good questions*)
* *Extras* – Did the presenters do something “extra” that caught the attention of the audience or kept the audience entertained and interested?

**Other Resources**

You already know how to use SAS EM. You can also use any of these packages, as well (or any software at all).

* [Weka](http://www.cs.waikato.ac.nz/ml/weka/) Data Mining Software
* [R](http://www.r-project.org/) Statistical Software
* Stata, EViews, Minitab, NLogit, Matlab.

I am happy to help you in any way that I can. You can also reach TA for help regarding the use of software or database management. Good luck!