

E370 Fall Semester, 2009-10 Team Project – Component One – Proposal

Due Date: For Penalty-Free Feedback from your coach: YOUR Lab, October 8 or 9, 2009

For Final Grading: YOUR Lab, October 22 or 23, 2009

Due Time: Beginning of YOUR Lab

Due To: Your AI

Eighteen percent of your final grade depends on a team research project which spans semester concepts. The project will ask your team to use virtually all methods learned over the semester to describe and analyze a data set selected by your team, and to statistically describe the relationships that exist amongst and between the variables in that data set. The project has four components that are turned in over the semester. The components are 1) Proposal, 2) Description, 3) Inference and 4) Oral report. Following are the guidelines for the first component, Proposal.

Goals: Propose a plan for predicting or explaining a variable of interest to your team chosen from one of the data sets provided on the Team Project page of the course web site based on other variables also found in that data set. This section of the project is valued at 30 points. Do not allow the relatively small number of points convince you that this is an unimportant part of the project.

Tasks:

1. Become acquainted with the data sets that are posted on the Team Project page. Most of the data files are accompanied by a “code book,” a document that provides information about the meaning of the variables. Some of the data files have variable definitions included in them.
2. In your team, discuss and select a question you are interested in answering. Suppose there were a data set called “Fuel” which contained information about price of gasoline and other related variables. You might develop a question such as, “Why is there such huge variation in the price of gasoline in Bloomington compared to other nearby Indiana cities?” Essentially, if I were working on this question, I want to be able to predict the price of a gallon of gasoline. I might think it has something to do with day of the week, distance from nearest competitor, distance from nearest supplier, whether the station was in Bloomington or some other city, the population within a specified area around the station, the income of the population within a specified area around the station, the University calendar, the weather, the level of international tension, and perhaps some other things.
3. Select relevant variables from the data set you have chosen to use.
 - a. You must have one variable you want to explain (dependent). The dependent variable (for example, the price of a gallon of gasoline) must be a numerical variable.
 - b. You want also to have several variables that you think might explain (independent) your dependent variable. You will be using Excel to do the statistical work, so bear in mind that Excel cannot work with more than 16 independent variables in the most sophisticated technique we will learn. The independent variables must include
 - i. at least one categorical variable (for example, Located in Bloomington or not, above) and
 - ii. at least one quantitative variable (for example, distance from nearest competitor, above).
 - iii. Selecting only two independent variables will be interpreted as a failure to act in good faith and points will be deducted. Select as many variables as needed to fully explain your dependent variable.

- c. Be certain that each observation (each row) has complete data for each variable you have chosen. If it does not, you must “clean” the data by removing any observations (this would be a whole row) where information is incomplete.
4. State the question you are answering. Explain carefully why the question is of interest to you and your team. Explain why your data will enable you to answer the question.
5. Provide a complete list of your variables including definitions, carefully and succinctly describe each variable in words and discuss the relationship you think may exist between the variables. Be sure to say how each variable is measured including the units it is in.
6. For penalty-free feedback, turn a draft of your proposal in to your coach during your lab October 8 or 9, 2009. **A Team Contact Person must be clearly indicated**, along with his/her email address. Your coach will provide feedback **within one week** of your turning in your proposal, by emailing the Team Contact Person.
7. Provide complete documentation for the source of your data. You should be able to find this in the code book. If you are unable to find the source of your data in the documentation, contact me.
8. Should you miss the final grading deadline for this section, you will forfeit 100% of the points for this section.

Document Format:

1. The document you will turn in to your coach during your lab October 8 or 9, 2009 is your Project Proposal and Data.
2. The proposal must be word-processed with one inch margins on all pages.
3. The proposal may not be longer than 400 words.
4. The text must be double spaced and printed in black ink.
5. The font must be 12 point Times New Roman in size.
6. Each paragraph must be indented.
7. At the top of the first page must be a title of your proposal, the date and a list of the members of your team.
8. An example of a possible first page follows the description of the data sets.

A Brief Description of the Data Sets

CARS 2004	Specifications are given for 428 new vehicles for the 2004 year. The variables recorded include price, measurements relating to the size of the vehicle, and fuel efficiency.
Education	An international data set from a UNESCO survey about educational spending and student enrollments, in addition to some other demographic variables are presented here.
Fringe	The data are at the wage-earner level and include information about hourly wage; personal, marital and household demographics; work experience; location; union affiliation; industry type and value of fringe benefits associated with that persons work.
Housing Survey	The information in this data set includes household energy costs, residence specific information, labor force participation and household income from a sample of Indiana homes.
Major League Salaries	The data are a set of Major League Baseball players who played at least one game in both the 1991 and 1992 seasons, excluding pitchers. This data set contains 1992 salaries along with performance measures for each player from 1991.

Manufacture The data set is information about the manufacturing sector of Ecuador and include some very high level factors affecting market quota, vertical integration and productivity. The data are from a paper written by one of our AIs, which adds a certain measure of interest to the data.

Used Cars For this data set, a representative sample of over eight hundred, 2005 GM cars were selected, then an algorithm was developed following the 2005 Central Edition of the Kelly Blue Book to estimate retail price.

World Bank This data set includes a variety of information for 112 countries including such things as fertility rates, mortality rates, CO2 emissions, power consumption, immunization rates and internet usage rates.

Additional Data Option:

It is possible that you may have a data set that you would like to use or are able to locate such a data set in a timely fashion. This semester you will be allowed to use your own data set, as long as it conforms to the following rules.

1. The data may not come from any survey you write and perform.
2. The data may not be any data used in a previous semester.
3. The entire data set must be made available to your lab coach at the time the proposal draft is due.
4. The data set must be characterized by
 - a. at least 100 good, complete observations.
 - b. The data must be real data.
 - c. The dependent variable MUST be numerical.
 - d. There must be at least two independent variables which are numerical.
 - e. There must be at least two independent variables which are categorical.
 - f. There must be more than four independent variables.
 - g. The data set must be accompanied by a "code book" which includes
 - i. A definition of each variable
 - ii. There must be sufficient detail in the definition that units are clear.
 - iii. The code book must include details about the source of the data.

A Possible Top of the Front Page

Team 20002:

Sleepy

Sneezy

Dopey

Doc (Team Contact Person: email: doc@indiana.edu)

October 8, 2009

A Sample First Page (this is the title)

Body of proposal and introduction of data.