UNIT EIGHT: DATA ANALYSIS PROJECT

All Excel output should be copied into a single Word document where you must enter all of your responses to the questions below. Format the document professionally so it flows well. Include a table of contents.

- Choose any published database from the internet or Bethel library (such as those from the Census Bureau or any financial sites). You may opt to use one of the data files provided by the instructor if applicable.
- Get advanced approval from the instructor on your chosen database.
- If the file is large, randomly choose 200 of the observations from the data.
- Explain each variable in the file that you are analyzing. Be sure your file includes at least 3 scale variables and at least 2 nominal variables.
- Conduct a descriptive analysis on any 2 interval / ratio variables you wish using *Descriptive_Statistics.xls* and *Frequency_Distribution.xls*. Explain the output.
- Conduct 3 different hypothesis tests of your choice using appropriate variables from the file (note: you must use 3 different tests and not run one test on 3 different variables). In each case, state the variables being tested as well as the hypothesis, decision and conclusion. Use 3 of the following (*1-Sample Test for Means, 1-Sample Test for Proportions, 2-Sample Test for Means – Independent Samples, 2-Sample Test for Means – Paired Samples, 2-Sample Test for Proportions, Analysis of Variance, Chi Square Goodness of Fit Test, Chi Square Test of Independence, Correlation Test).*
- Develop a model to predict an interval / ratio variable using at least 2 other variables. Use *Multiple_Regression.xls* and state the regression model and which variables are or are not significant. Also, use the model to make a prediction by making up values for each of the independent variables.
- Write a one to two page summary of your findings. Include the data file in the appendix.

The project is due at the close of Week 8. **You may work alone or in a team of 2** (you choose your own partners and both of you must let your instructor know of your intent to work together).

You may use a data set from the internet or from your workplace, or you may use one of the files provided on www.drjimmirabella.com/bethel The files are described here, and the variables are described within the files. If there is anything confusing about these data files, please ask your instructor.

BASEBALL: This file includes actual team by team data for the 1997 MLB season. The key variable to predict in Multiple Regression Analysis is the number of wins (or possibly the attendance). Lots of interesting analysis possibilities here, including how team salary relates to a team making the playoffs, or whether money buys wins, or how wins relate to attendance, or how performance on the field relates to the field surface, etc. If you know something about baseball, this file should make sense to you.

CARS: This file is self-explanatory after you open it. Several variables describe the car (sports car, SUV, engine size, horsepower, etc.), and several describe the car's performance (CityMPG and Highway MPG). It also includes the Dealer Cost and Suggested Retail Price. The key variable to predict in Multiple Regression Analysis is the Suggested Retail Price. Lots of crosstabulation options for Chi Square Analysis, lots of ANOVA and t-test options in which you analyze miles per gallon or price as a function of any of the many variables included.

LOW BIRTH WEIGHT: This file looks at factors that might predict a baby being born with low birth weight. Birth weights of 5.5 pounds or less are considered low in this file. Use the actual birth weight as the key predicted variable in Multiple Regression Analysis. Lots of variables about the mother regarding her weight, race, medical problems, and doctor's visits can be used for Chi Square analysis or as factors in ANOVA's or t-tests.

MUTUAL FUNDS: This file looks at Large Cap, Mid Cap and Small Cap funds with either Growth or Value objectives. Some funds have fees. Funds are either high, average or low risk. Assets range from 50.7 million dollars to 66.5 billion dollars. For Multiple Regression Analysis, you can choose to predict any of the three Return rates (measured in percents). Lots of categorical variables to choose from in a Chi Square Analysis or as factors to analyze differences in mean return rates.

TIPS: This file includes data on 75 patrons at the Spaghetti Warehouse on a given day. The key variable here is the Tip Rate or the Tip Total. If you wait tables there, under what circumstances are you most likely to get a better tip? You can compute mean Bills or Tips or Tip Rates as a function of the meal time, the party size or the size of the party at the table.

Note that you should not use a nominal variable with 3 or more values in the Multiple Regression Analysis (unless you convert to dummy variables, but that is unnecessary here).