Stat 666
Mini-Project #2
Due TBA

To be done in groups of two...choose your own partner then send me an email to let me know the name of the person with whom you're working.

On the course webpage, you will find the files `oliver2a` and `oliver4a`. Each file contains observations on 8 fatty acids found in olives in two agricultural regions: Region 2 and Region 4. The acids measured are listed in the first line of each file. You may treat the observations (rows) in each file as a random sample of olives from the region of interest, with the two samples being drawn independently by different organizations with potentially different data collection procedures, chemical analysis tools, and data censoring mechanisms.

There are two basic dimensions of the analysis required here.

1. Assess whether the region 2 sample for the given year deviates from the historical region 2 average of $\boldsymbol{\mu_{02}} = (1300, 120, 265, 7310, 820, 45, 65, 28)'$, where the eight components refer to palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, and eicosenoic acids, respectively. IF there is a significant difference, comment on the nature of the difference between the sample averages and the hypothesized means. Specifically, which of the 8 variables are important when comparing to the hypothesized mean vectors. Note that because of missing values present in the data, you will need to make use of the EM algorithm and multiple imputation in the multivariate hypothesis testing setting. Is there any reason to suspect that the "missing at random" assumption has been violated? (This assessment of random missingness will require some creativity. Bonus points to students that excel in addressing this knotty issue. Hint: look for any statistically significant tendency for partial censoring in some observations.) Repeat the analysis for region 4, comparing the region 4 data with the historical region 4 average of $\boldsymbol{\mu_{04}} = (1230, 105, 275, 7360, 830, 41, 75, 38)'$.

2. Because these agricultural regions are adjacent to each other, and because of modern agricultural procedures, an agronomist believes that region 2 and region 4 olives have evolved to have essentially the same profile in terms of the eight fatty acids. Test whether or not this claim is valid. Comment on the nature of the difference between the two sample averages. Specifically, which of the 8 variables are important when comparing the means for the two groups. Additionally, address the following question of interest to the client: can linoleic and linolenic acids be dropped from the list of fatty acids without significantly decreasing the separation of the two samples? (I.e., for separating the two groups, do linoleic and linolenic acids contribute anything significant beyond the information already available in the other 6 acids?)

You should turn in a write-up of no longer than **5 pages** (11 pt font, 1 inch margins, single spaced). This includes all tables and figures. Your reasonably-well-commented computer code (R, SAS, or C) for your EM function should be attached to the back of your report in an Appendix. Your EM function should be a function that takes a matrix containing missing

values, and returns a matrix with new values imputed. Also include in the Appendix the updated matrix for region 2 and region 4 after running your EM function. (The Appendix with code and printouts of your matrices does not count against your page limit.) You may assume that your client has an MBA and a basic understanding of statistical methods (anova, regression, hypothesis testing, confidence intervals, etc.), but the reader is NOT a professional statistician. However, your client employs an M.S. level statistician that will also be reading the report to check for accuracy. Hence, a careful explanation of the problem, techniques, and results are crucial. Keep in mind that I am still another reader of the report—you need to convince me that you understand the principles discussed in class and in the text.

You are not allowed to discuss the results of your analysis with other members of the class (except for your project partner). You may, however, discuss the project with me.

Final notes:

Normality and outlier-detection analyses are a valuable part of any project. However, in the interest of keeping the project scope at a manageable level, for this assignment you may begin with data analysis directly.

And, as always, do not just flood the reader with a multitude of graphs and numbers, many of which are redundant. The key to a good project is carefully selecting which graphs and measures best support your claims.

Because of the amount of work required in this project, this project will get 1.5 times the weight of the first project.