CSC 424/334 Data analysis and Regression
Project description and guidelines

The project consists in the statistical analysis of a large data set using the techniques covered in class and other multivariate analysis techniques that the group may wish to investigate. Students are encouraged to suggest datasets for the project. There are several repositories for large datasets that are available online, and some students may have access to datasets themselves. If you do bring a dataset to the project, however, you must make sure you have the right to use the data! Students that want to suggest their own dataset must send me a request with the description of the dataset before starting the analysis.

The minimal requirement is that the dataset must contain at least 15 variables with a mix of at least half metric and ordinal variables with the rest being categorical or binary variables (a good mix of at least several of each type is encouraged to give you a range of options) and 300 or more observations. If your data set doesn't quite meet this, come see me and we can talk about it, but too many categorical variables gets difficult for applying many of the techniques from the course and too few samples can make it difficult to achieve stability on some of the techniques we will cover.

Project goals and report
The goal of this project is to analyze a large dataset with at least three different multivariate techniques (groups with fewer than 3 people will be allowed 2), or distinctly different approaches to using the same technique (your group should still use at least 2 distinct techniques in this case), to draw out interesting relationships that will either allow you to predict continuous values for new data, classify new data for categorical variables, determine probabilities of classification for binary variables, or determine the relationships between discrete properties of the data (what features tend to group with what features. Your group should explore both relational (regression, PCA, canonical correlation, etc.) techniques as well as classification methods (linear discriminant analysis, cluster analysis, multidimensional scaling, etc.) to analyze their set.

Each technique should be reinforced by appropriate statistical measures of fit, and cross-validation should be used when possible. Especially if you have a large sample set for your data, consider dividing your data randomly into a training set (to build the model) and a validation set (to test the model). This is particularly important for regression and discriminant analysis techniques. Also, you should employ leave-out-one or k-fold cross validation when appropriate.

You may use any software you wish to complete the analysis, but each group member must contribute to the set of analyses and should have their own SAS/R/Python/SPSS, etc. code to exhibit in the end for their individual report. The group will synthesize all the analyses completed by the group into a single report describing the main lines of analysis conducted by the group and the conclusions that the analyses drew. Group members not participating fully in the group work may receive less credit than other members of the group.

The analysis should be submitted as a report consisting of two sections:

1. **A non-technical summary** not longer than two typewritten pages, describing the conclusions of your statistical analysis to a non-technical audience. It should be intelligible to a person who does not know regression analysis. Suppose you are talking to your boss who does not know statistics or to a friend who is not familiar with statistical terminology. Include at least two graphs that indicate the most clearly some aspect of your conclusions.

2. **A technical summary** not exceeding 7 pages, with details of your statistical analysis. This section is intended for a statistically literate audience and must be written in a clear organized fashion.

For instance, you can organize the report into subsections such as

a. Exploratory analysis of the data.
b. Model selection.
c. Residual analysis and model diagnostics.
d. Analysis of the results and discussion.

3. Each team member will submit a 1-2 page summary of their work on the project to be included in an appendix of the main document. This should detail what role they played as a part of the team and what specific analyses or visualizations they performed.

4. **A brief evaluation of the work of each group member will be submitted separately by each group member as an appendix to the main document. These evaluations will form a portion of your final project grade.**

Your analysis should address the following points:

1. The exploratory analyses that you performed on the data. What groups of variables your group chose to analyze and what specific parameters of interest arose. Were there any nonlinearities that you had to account for?
2. What multivariate techniques you attempted and why these techniques were appropriate for your data. Since part of the requirement is that you explicitly explore at least three different techniques, your write-up should **clearly** indicate the different approaches that you attempted.
3. What kinds of variable selection methods did you use for building models? Did you explore automatic and/or manual selection methods? Did you explore PCA/Canonical Correlation for detecting underlying factors to relate?
4. When performing regression, you should be checking for multicollinearity among the independent variables, and exploring different ways of dealing with it (you should try different methods and see how the results compare)
5. When performing PCA/CFA you should detail how you chose the appropriate number of factors and what techniques you used to aid you in interpreting those factors (factor rotation, factor representation methods, etc.)
6. What kinds of categorical information was explored and how was it used in the analysis? Were you able to find statistical methods to classify samples according to categories (as in LDA), or were you able to detect grouping patterns for categorical properties?
7. FOR GRAUDUATE STUDENTS (e.c. for undergrads): Your write-up should also consider the practical significance of the results as well as the statistical significance.
8. FOR GRADUATE STUDENTS (e.c. for undergrads): Apply cross-validation techniques where appropriate to evaluate the predictive power of your model. Apply k-fold cross validation to check the adequacy of the model for out-of-sample predictions/classifications. If your analysis suggests two alternative models, you can use cross-validation to compare their predictive power and select the model with the lowest RMSE/R^2/etc.

You should integrate appropriate output and graphs from your statistical analysis into your document in a presentable format. Further details should be relegated to an **appendix** that may contain the SAS/R code, some graphs, computer output or supplementary information about the field of study. Graphs are a very important part of a complete analysis, so spend some good time thinking about them and construct them with care. Remember, you go through multiple graphs for test in a paper, you should do the same with graphs.

Clarity and synthesis in the report are important and will be rewarded. It is essential to be able to

communicate your analysis to other people. You must explicitly quote any result you will be using in your analysis.

Hints for the statistical analysis:
It is possible that you may not find a satisfactory model that fits adequately your data. Sometimes a data set may admit more than one satisfactory answer; sometimes there may be none. If the statistical analysis shows a lack of good models/fits/etc. for your data set, then you need to extensively detail what approaches you tried, what was unsatisfactory about them, and what remedies you attempted. If there is more than one suitable model, mention the pros and cons and compare their performance in predicting the response variable.

The final aim of any statistical analysis is the understanding of a phenomenon or the investigation of a scientific problem, which your data arise from. Remember that analysis techniques we covered are mathematical representations of such a problem and the interpretation of the parameters values will give you insights about the relationships of the variables in the problem.

CSC423 Final project Grading Rubric

Group Report	Excellent	Good/Fair	Poor
Layout and clarity	Report is clear and neatly organized, with appropriate use of headings and tables to enhance readability. Report is well written with appropriate use of grammar and statistical terminology. (5pts)	Report is mostly clear with parts that are not well organized under sections. Use of appropriate statistical terminology is limited. (3pts)	Report is not clear. The layout is cluttered and not organized in sections. Major editing and revision are required. Errors in spellings, capitalization, punctuation and grammar distract readers. (0-1 pts)
Non-technical summary	Summary describes the conclusions clearly, concisely and is written in a language that is appropriate for a non-technical audience (5pts)	Summary describes the conclusions clearly, concisely but contains statistical jargon that is inappropriate to a non-technical audience (3 pts)	Summary is unclear and uses technical terms and language that is inappropriate to a non-technical audience (0-1 pts)
Exploratory data analysis	The basic variables were well explored for their distributions, and correlations. Metric parameters were explored for their dependence on categorical parameters. (4 pts)	The analysis of the distributions, correlations, or dependence between metric and categorical variables is incomplete or is incorrectly interpreted. (2-3 pts)	The analysis of the distributions, correlations, or dependence between metric and categorical variables is missing or contains major flaws. (0-1 pts)
Exploratory Graphs	The basic exploratory analysis is augmented by appropriate summary graphs which are correctly produced and interpreted (3 pts)	The exploratory graphs are incomplete or contain minor flaws (2 pts)	Exploratory graphs are missing or contain major flaws (0-1 pts)
Approach 1 Execution	The first approach you chose is appropriate for the data chosen and is properly executed. (5 pts)	The first approach you chose is not completely suited to the data or has minor problems in its execution. (3 pts)	The first approach you chose is inappropriate for the data or has major problems in its execution. (0-1 pts)
Approach 1 Analysis	Approach assumptions are checked via appropriate statistical measures and/or	Model assumptions are checked but some steps in computing or interpreting statistical	Major flaws in analysis, missing fitness missing analysis of outliers and

	residual analysis. Fitness measures are correctly computed and interpreted. (5 pts)	measures of fitness or residuals are missing or incorrect. (3 pts)	influential points (0-1 pts)
Approach 2 Execution	The second approach is appropriate for the data chosen, is substantially different from the first approach and is properly executed. (5 pts)	The second approach is similar to the first approach, is not completely suited to the data or has minor problems in its execution. (3 pts)	The second approach you chose is inappropriate for the data or has major problems in its execution. (0-1 pts)
Approach 2 Analysis	Approach assumptions are checked via appropriate statistical measures and/or residual analysis. Fitness measures are correctly computed and interpreted. (5 pts)	Model assumptions are checked but some steps in computing or interpreting statistical measures of fitness or residuals are missing or incorrect. Incomplete analysis of outliers and influential points (3 pts)	Major flaws in analysis, missing fitness missing analysis of outliers and influential points (0-1 pts)
Approach 3 Execution	The third approach is appropriate for the data chosen, is substantially different from the first two approaches and is properly executed. (5 pts)	The third approach is similar to the first two approaches, is not completely suited to the data or has minor problems in its execution. (3 pts)	The third approach you chose is inappropriate for the data or has major problems in its execution. (0-1 pts)
Approach 3 Analysis	Approach assumptions are checked via appropriate statistical measures and/or residual analysis. Fitness measures are correctly computed and interpreted. (5 pts)	Model assumptions are checked but some steps in computing or interpreting statistical measures of fitness or residuals are missing or incorrect. Incomplete analysis of outliers and influential points (3 pts)	Major flaws in analysis, missing fitness missing analysis of outliers and influential points (0-1 pts)
Conclusions	Results from the three approaches results are correctly interpreted and conclusions are clearly reported. Including practical significance (5 pts)	Interpretation of results contains one or two errors. (3 pts)	Interpretation of results contains three or more errors, or is missing. (0-1 pts)

Appendix	Well organized with full code/scripts included– graphs and output are labeled appropriately. (4pts)	Not well organized – graphs and output have missing labels. Not all code included (2pts)	Cluttered and confusing. Many missing elements (0-1 pts)
Punch list addressed	Items in the punchlist have been specifically addressed in the writeup and the issues raised in the punchlist have been corrected (4-5pts)	An element from the punchlist was not addressed or were inadequately addressed (2-3 pts)	Many or all elements of the punchlist were not addressed or were inadequately addressed (0-1 pts)
Group members	Information received by due date (0 pts)	Information not received by due date (-10 pts)	
Extra Credit for technique not covered in class	Analysis includes use of multivariate technique beyond those covered in class (+ 5 pts)	No additional technique covered (+0 pts)	
Total	**/ 60 pts**	+ 5 points for possible extra credit	

Individual Report	Excellent	Good/Fair	Poor
Layout and Clarity	Report is clear and neatly organized, with appropriate use of headings and tables to enhance readability. Report is well written with appropriate use of grammar and statistical terminology. (10 pts)	Report is mostly clear with parts that are not well organized under sections. Use of appropriate statistical terminology is limited. (6 pts)	Report is missing or unclear. The layout is cluttered and not organized in sections. Major editing and revision are required. Errors in spellings, capitalization, punctuation and grammar distract readers. (2-0 pts)
Summary of Work	Summary describes your contributions clearly and concisely. Report shows evidence of clear and equitable contribution to the group's effort. Effort is described in language that shows understanding and proper usage of the statistical methods applied. (10-8pts)	Parts of summary are unclear, report shows limited areas of misunderstanding or improper usage of statistical techniques, or report shows some lack of effort. (6 pts)	Summary of work is missing, shows lack of effort, or summary shows misunder-standing of statistical techniques. (2-0 pts)
Summary of Takeaways	Report includes a clear summary, with correctly interpreted concepts, of what you learned about multivariate statistics from the project. (5 pts)	Report of what you learned lacks some clarity or is unspecific concerning what you learned about multivariate statistics from the project. (3 pts)	Summary of takeaways is missing or lacks a clear description of what you learned about multivariate statistics from the project. (2-0 pts)
Evaluation	Submitted (5 pts)	Not submitted (0 pts)	
Partner's Evals	Grade depends on their evaluation (10-0 pts)		
Total	/ 40 points		

Total Project Grade: / 100 pts