Applied Integrative Projects in Data Analytics I
[Introduction to SAS and Big Data]
(DA520, CRN: xxx*, Fall2017)*

Instructor:	Paul Yan
Contact Information:	yany@canisius.edu, (716) 888-2604
Lecture:	In a computer lab
Office hours:	TBA
Prerequisite:	DA500
Language	SAS
Capacity	Since hands-on is so important, the maximum number of students is 20.
Textbooks:	Delwiche, Lora D. and Susan J. Slaughter, Little SAS book: A primer, SAS Institute Inc., 5th ed, https://www.sas.com/storefront/aux/en/splsb/65423_excerpt.pdf Step-by-Step Programming with Base SAS® Software https://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_step_10071.pdf
Websites:	*SAS support:* http://support.sas.com/downloads/index.htm SAS Press and SAS Documentation Example Code and Data: *http://support.sas.com/documentation/onlinedoc/code.samples.html* My website: http://www3.canisius.edu/~yany/sas.shtml
QR codes	
One-line R codes	*> source("http://canisius.edu/~yany/da520.R")* Note: I will explain this line in week 2
Capacity	Since hands-on is so important, the maximum number of students is 15.
Expected output	Here are several examples students will be able to at the end of this course. 1) able to process DTAQ extra-high frequency data (see Appendix A. 2) able to process US Census data (10G) 3) able to process stock data from 1926 to 2015 about 5G 4) able to process company's financial statements from 1950 (about 30G)
Course Description:	In this course, students would learn SAS. Since the focus is on hands-on, all lectures would be conducted in a computer lab. Students learn how to input various types of data into SAS, such as text, csv, binary and sas7bdat. How to clean data is an important skill

	students are expected to master. Students learn how to deal with missing variables and run basic sample statistics such as mean, standard deviation, minimum and maximum. Many visualization techniques would be taught. In addition, students learn how to run some basic statistical functions, such as linear regression. Since this course is a preparation for the next course titled "Applied Integrative Projects in Data Analytics II", students could start to think about their next big projects. [Note: the instructor would distribute, after the mid-term, a list of potential big projects related to various domains, such as Economics, Accounting, Finance, Marketing and Health Care.]
Four objectives:	1) learn SAS, 2) hands-on experience, 3) use tons of real-world data , 4) get familiar with a few potential projects which students might choose from as their project for DA521. More detail: create and run SAS programs in a PC environment; read raw input files in various formats, such as text, CSV and sas7dat; create SAS datasets; create new variables from other data; use basic SAS procedures to summarize a give data set numerically and graphically; annotate SAS output with titles, labels, and formats; work with SAS datasets: sort, subset, merge, and re-format; use SAS procedures for basic statistical analysis, conduct T-test, F-test to test equal variance, equal means; export SAS data and output for further analysis by other software, such as R and Excel.
Academic Integrity:	Students are expected to know and understand college policies with regard to Academic Integrity Code. Violations of academic integrity will be prosecuted fully. Please note that you are responsible for reporting any instances where other students have violated these policies. Failure to do so will result in penalties as well. If you have any questions about this policy, please see the instructor.
Attendance Policy:	Attending classes regularly is required. Before-class preparation and in-class participation is an integral part of this course. Students are strongly encouraged to participate in class discussions and ask questions. Students are encouraged to discuss current events relevant to this course or their own experiences. Homework problems are regularly assigned.
Academic and Accessibility Support Services:	The GRIFF Center for Academic Engagement provides comprehensive programs, tutoring services, and resources to support student academic and career success. If you would like to learn more about academic support, please stop in Old Main 013 or call 716-888-2170. Visit the GRIFF Center webpage at: http://www.canisius.edu/griff-center/ . Accessibility Support (716-888-2170), which is located in the Griff Center for Academic Engagement (OM 013), is responsible for arranging appropriate academic accommodations for students with documented disabilities. If anyone in this course falls into this category, please contact Accessibility Support so that an appropriate course of action may be determined. For additional information, see http://www.canisius.edu/dss/
Grade Evaluation:	HW (about 9) 25% Data Cases (3) 20% Mid-term 20% Final 25% Class participation 10% ---------- -------- Total 100%
Teaching Methods:	Each class will be consist of two parts: lecture (including discussion of homework) and hands-on.

Course Schedule:	For the detailed schedule, see below. I reserve the right to change the course schedule throughout the semester. Changes to the schedule will be announced in class or via email.

Tentative schedule

Week	Description	HW
1	Self introduction, discuss the objectives of this course, SAS language, syllabus etc. Why SAS? SAS vs. R [1] Chapter 1: Getting Started using SAS Software 1 – Introduction to SAS - SAS environment, program syntax, structure of data, types of data	
2	Chapter 1: (continued) - read SAS log - reading in and displaying data - running program, generating log and output	HW1
3	Chapter 2:Getting your Data Into the SAS Reading in Data - list input, comma and tab delimited data, data from Excel, column input, informat - reading data from an external file - problems in reading data	HW2 Data case #1
4	Chapter 3: Working with your data describing Data I - PROC PRINT, MEANS, UNIVARIATE, SPLOT - summary statistics, graphical displays, controlling the output	HW3
5	Chapter 4: Sorting, Printing and Summarizing Your Data - PROC FREQ, SPLOT, CORR,REG - frequency distribution, 2-way, tables, correlation, simple regression	HW4
6	**Mid-term**	
7	Chapter 5: Enhancing Your Output with ODS – Creating Variables in the Data Step, direct assignments, if, then, else statements, SAS functions, handling missing data	HW5 Data Case #2
8	Chapter 6: Modifying and Combining SAS data sets – Formatting Output; Working with Dates - titles and labels - PROC FORMAT - FORMAT statement - Working with dates	HW6

[1] https://thomaswdinsmore.com/2014/12/01/sas-versus-r-part-1/
https://thomaswdinsmore.com/2014/12/15/sas-versus-r-part-two/

Continued

Week	Description	HW
9	Chapter 7: Writing Flexible Code with the SAS Macro Facility - sub-setting and merging datasets - SET and MERGE statements, KEEP option and statement - WHERE statement and logical if statements - creating and using permanent datasets - LIBNAME statement	HW7
10	Chapter 8: Visualize Your Data ODS graphs, - ODS OUTPUT, OUTPUT statement - PROC RANK	HW8 Data Case #3
11	Chapter 9 : Using Basic Statistical Procedures - Chi-square, T-Tests, ANOVA, Non-parametric tests, reading in frequency counts as raw data PROC FREQ, TTEST, ANOVA, GLM, NPAR1WAY - modeling binary data, logistic regression - modeling continuous data, linear regression - PROC LOGIST, REG,LIFETEST, PHREG	HW9
12	Chapter 10: Exporting Your Data Generate a text, csv file PROC EXPORT, TABULATE	
13	Chapter 11: Debugging your SAS programs Read SAS log , Print g few lines Proc contents data=temp;run;	
14	**Final Exam**	

List of potential data cases:

#	Description
1	Which party, Republican or Democratic, would manage the economy better?
2	Generate 20 SAS data sets from Prof. French's Data Library[2]
3	Generate a dozen SAS data sets from Federal Reserve Bank's Data Library[3]
4	Estimate VaR for two dozen stocks[4]
5	Market risk (beta) estimation for a dozen stocks[5]
6	Replicate the S&P500 index

[2] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[3] http://www.federalreserve.gov/econresdata/default.htm
[4] http://finance.yahoo.com/
[5] http://finance.yahoo.com/

Appendix A: Size of TDAQ data sets

Index of /Historical Data Samples/Daily TAQ/

Name	Size	Date Modified
[parent directory]		
EQY_US_ALL_BBO_20031203.zip	366 MB	3/15/16, 1:33:00 PM
EQY_US_ALL_BBO_20031204.zip	382 MB	3/14/16, 2:19:00 PM
EQY_US_ALL_BBO_20131218.zip	6.4 GB	1/28/14, 12:00:00 AM
EQY_US_ALL_BBO_20141030.zip	6.6 GB	11/11/14, 12:00:00 AM
EQY_US_ALL_BBO_20150805.zip	391 MB	9/16/15, 12:00:00 AM
EQY_US_ALL_BBO_20160627_prod.gz	6.3 MB	7/7/16, 11:02:00 AM
EQY_US_ALL_BBO_ADMIN_20150805.csv.zip	66.9 MB	8/24/15, 12:00:00 AM
EQY_US_ALL_NBBO_20131218.zip	2.0 GB	1/28/14, 12:00:00 AM
EQY_US_ALL_NBBO_20150805.zip	3.0 GB	8/24/15, 12:00:00 AM
EQY_US_ALL_NBBO_20160627_prod.gz	1.3 MB	7/7/16, 2:56:00 PM
EQY_US_ALL_REF_MASTER_20131218.zip	357 kB	1/28/14, 12:00:00 AM
EQY_US_ALL_REF_MASTER_20160111.zip	374 kB	3/15/16, 4:02:00 PM
EQY_US_ALL_REF_MASTER_20160112.zip	373 kB	3/15/16, 4:02:00 PM
EQY_US_ALL_REF_MASTER_PD_20160111.txt	812 kB	3/15/16, 4:02:00 PM
EQY_US_ALL_REF_MASTER_PD_20160111.xls	2.7 MB	3/15/16, 4:02:00 PM
EQY_US_ALL_REF_MASTER_PD_20160112.txt	812 kB	3/15/16, 4:02:00 PM
EQY_US_ALL_REF_MASTER_PD_20160112.xls	2.7 MB	3/15/16, 4:02:00 PM
EQY_US_ALL_TRADE_20031203.zip	58.2 MB	3/15/16, 1:30:00 PM
EQY_US_ALL_TRADE_20031204.zip	59.6 MB	3/14/16, 2:20:00 PM
EQY_US_ALL_TRADE_20131218.zip	298 MB	1/28/14, 12:00:00 AM
EQY_US_ALL_TRADE_20141030.zip	271 MB	11/11/14, 12:00:00 AM
EQY_US_ALL_TRADE_20150805.zip	654 MB	9/16/15, 12:00:00 AM
EQY_US_ALL_TRADE_ADMIN_20150805.csv.zip	69.8 MB	8/24/15, 12:00:00 AM

Source of the above data sets: ftp://ftp.nyxdata.com/Historical%20Data%20Samples/