Applied Integrative Projects in Data Analytics I
[Introduction to SAS and Big Data]
(DAT521, CRN: 38489, Fall2019)

Lecture: August 26, 2019 - December 07, 2019

Instructor:	Paul Yan
Contact Information:	yany@canisius.edu, (716) 888-2604
Lecture:	TTR: 4:00pm -- 5:15pm @ SH 036 (SH stands for Science Hall)
Office hours:	Tuesdays: 2:50m – 3:40pm (CT308)
Prerequisite:	DAT500
Language	SAS
Capacity	Since hands-on is so important, the maximum number of students is 20.
Textbook	My lecture notes (Finanical Data Analytics using SAS)
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://datayyy.com/abc.R")* *or* *source("http://canisius.edu/~yany/abc.R")* Note: I will explain the above lines of R code when we meet.
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.]
Program objectives	Objective A: Domain Knowledge: Students will be able to apply the computational and statistical methods and analytical tools to strategic and tactical decision making for at least one domain area. In business, for example, this might be: accounting, economics, finance, management, or marketing. Objective B: Adaptable grounding in applied statistics. Students will be able to use the basic principles of probability theory in a variety of contexts, including both classical statistical approaches and computational based methods. Students will be familiar with one modern statistical software platform and will be able to readily adapt to others. Objective C: Flexible computational skills. Students will have a strong working knowledge of at least one general purpose programming language, and will be able to work with a range of data structures within those languages. Students will also be familiar with databases and the programming techniques needed to work with Big Data.
Course 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 DAT522. 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:	Data Cases (4-5) 20% Mid-term 20% Final 20% Term project 20% Term project presentation 10% Class participation 10%
Teaching Method:	Each class has two parts: lecture (including discussion of data cases) and hands-on.

Tentative schedule

Week	Date	Description	Data case
1	8/27 8/29	Self-introduction, discuss the objectives of this course, SAS language, syllabus etc. Chapter 1: Introduction to SAS - Compare SAS with R, Python, Matlab and C/C++ - PC SAS, - read SAS log, - reading in and displaying data	
2	9/3 9/5	Chapter 2: UNIX SAS SSH, Putty, UNIX text, editor: vi, SAS basic rules (conventions)	
3	9/10 9/12	Chapter 4:Data input and date variable - CSV, tab delimited data, Excel, SAS input, informat - multiple commas, missing data items - problems in reading data, PROC PRINT - LIBNAME statement, input SAS data sets - working with date variable	data case #1
4	9/17 9/19	. Chapter 5: Open data - Yahoo!Finance, Google finance, SEC filings, - UCI Machine Learning data depository, - Federal Reserve Bank's Data library, Census data	data case #2
5	9/24 9/26	Chapter 6: Data manipulations - missing data, informat, summary statistics - - PROC FREQ, proc datasets - sub-setting and merging datasets - sort and MERGE, KEEP=, drop= - WHERE and if conditions, daily to weekly, monthly etc	data case #3
6	10/1 10/3	Chapter 7: SAS PROC - Critical values for T-test, F-test - Proc TTEST, PROC PRINT, PROC means - Proc logic, proc means,Proc reg, pric univariate -PROC TABULATE	data Case #4
7	10/8 10/10	Chapter 8: Data output - writing csv, text, PROC EXPORT, SAS data sets - binary data set, a database format , OUTPUT statement	Data Case #5
8	**10/15** 10/17	**Mid-term** Chapter 9: SAS simple macro - Write Flexible Code with the SAS Macro - %let and call variable &, SAS macro loops - Call a macro within another macro loop - %include (call an outside SAS program)	

Week	Date	Description	Term-project
9	10/22 **10/24**	Chapter 9 (Continued) Chapter 30: Term projects - A list of topics and requirements for a term project **No Class (I will attend the FMA annual meeting)**	Discuss the list of term projects and one of term
10	10/29 10/31	Chapter 10: Simple graphs and data visualization - Proc PLOT, proc gplot - ODS graphs, - ODS OUTPUT, - PROC RANK	Discuss one term project
11	11/5 11/7	Chapter 11: String manipulations - $ for string variable, character variable verification - Length feature - How to move a string from a phrase - String to numeric ,Substr() function - Parsing a string, scan() - Tranwrd() , location of a substring in a long string	Discuss one term project
12	11/12 11/14	Chapter 12: Dealing with big data - High frequency data - SEC filings - Introduction to TAQ and TORQ - NYSE millisecond-by-millisecond sample data	Discuss one term project
13	11/19 11/21	Chapter 13: Text analysis - High frequency data - SEC filings - Introduction to TAQ and TORQ	
14	11/26 **11/28**	Chapter 15 : SAS Matrix Manipulation (IML) - Why matrix operation? - Enter row, column and matrix, retrieve SAS data into IML - Print, nrow(), ncol(), Solve multiple equations - Optimization procedure **Thanksgiving Holiday - No Classes**	
15	12/3 12/5	Student presentations Student presentations	
		Final Exam	

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/

References

50 Most Frequently Used UNIX / Linux Commands (With Examples), https://www.thegeekstuff.com/2010/11/50-linux-commands/?utm_source=feedburner

Basics of SAS programming Lauguage, https://www.slideshare.net/guest2160992/basics-of-sas-programming-language

Black, Fischer and Litterman, Robert (1992) "Global Portfolio Optimization.", *Financial AnalystsJournal*, Sept/Oct 1992.

Brandt, M.W., Santa-Clara, P. and R. Valkanov, 2009, Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns, *Review of Financial Studies*, 22, 3411-3447.

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

R or SAS: which one is the best statistical software used in medical field?, https://www.researchgate.net/post/R_or_SAS_which_one_is_the_best_statistical_software_used_in_medical_field

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

R vs SAS (Comparison and Opinion), http://www.learnanalytics.in/blog/?p=9

R vs SAS, why is SAS preferred by private companies?, http://stats.stackexchange.com/questions/33780/r-vs-sas-why-is-sas-preferred-by-private-companies

SAS versus R debate, http://insidebigdata.com/2014/03/01/sas-versus-r/

SAS vs. R (vs. Python) – which tool should I learn? https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

SAS vs. R (Part I), https://thomaswdinsmore.com/2014/12/01/sas-versus-r-part-1/

SAS vs. R (Part II), https://thomaswdinsmore.com/2014/12/15/sas-versus-r-part-two/

SAS Language, https://en.wikipedia.org/wiki/SAS_language

SAS cheatsheet, https://sites.ualberta.ca/~ahamann/teaching/renr480/SAS-Cheat.pdf

SAS cheatsheet (2 pages), https://exchange.canisius.edu/owa/auth/logon.aspx?replaceCurrent=1&url=https%3a%2f%2fexchange.canisius.edu%2fowa%2f

SAS Cheat Sheet for Stat-342, Carl James Schwarz, http://people.stat.sfu.ca/~cschwarz/Stat-342/Exams/sas-cheat-sheet.pdf

SASCheat, http://www.theprogrammerscabin.com/SASCheat.pdf

SSH download , https://www.ssh.com/ssh/download/

SAS Academic Resources, https://www.sas.com/en_us/learn/academic-programs/resources.html#

SAS dataset files (*.sas7bdat), http://www.principlesofeconometrics.com/sas.htm

Step-by-Step Programming with Base SAS® Software, https://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_step_10071.pdf

Twenty most commonly used Unix commands listed in order: http://www-users.york.ac.uk/~hcb1/unix20.html

Yan's SAS website, http://www3.canisius.edu/~yany/sas.shtml

Financial Data Analytics using SAS

Table of contents

```
> .all
function(){
"
 *----------------------------------------------------------------*
 * Financial Data Analytics using SAS                             *
 *----------------------------------------------------------------*
 * c1  Introduction to SAS          c16 Advanced macro            *
 * c2  UNIX SAS                      c17 MySQL,SAS,R, and Python   *
 * c3  Free SAS                      c18 Various distributions     *
 * c4  Data input                   c19 T-test, F-test, Nomality test *
 * c5  Open data                    c20 Linear models, VaR        *
 * c6  Data minipulation            c21 Time series               *
 * c7  SAS proc                      c22 Optimization              *
 * c8  Data output                  c23 Monte Carlo Simulation    *
 * c9  Simple macro language         c24 Portfolio theory          *
 * c10 Simple graphs                c25 Credit risk analysis      *
 * c11 Simple string manipulation   c26 Machine Learning          *
 * c12 Dealig with big data         c27 Seminar papers/replications (1) *
 * c13 Text analysis                c28 Seminar papers/replications (2) *
 * c14 Parsing SEC filings          c29 Seminar papers/replications (3) *
 * c15 SAS Matrix manipulation/IML  c30 Term projects             *
 *----------------------------------------------------------------*
```