Canisius COLLEGE
Where leaders are made

Applied Integrative Projects in Data Analytics II
[Machine learning, big data analytics using R and SAS]

(DA521, *CRN:* xxx, *Spring 2018)*

Instructor:	Paul Yan
Contact Information:	yany@canisius.edu, (716) 888-2604
Lecture:	In a computer lab
Office hours:	TBA
Prerequisites:	DA500 and DA520
Languages of the course	R (for simple and small data sets) SAS (for complex, such as mutiple data sets and we have to merge them; for big data sets such as 6.4G, see a typical example below ftp://ftp.nyxdata.com/Historical%20Data%20Samples/Daily%20TAQ/
Project languages	In addition to R and SAS, students could use Python, Matlab or Perl to finish their term projects. However, those three lanugages will not be taught in this course.[1]
Capacity	Since hands-on is so important, the maximum number of students is 20.
Textbooks and other teaching materials:	1) Delwiche, L. and S. Slaughter, Little SAS book, 5th ed. SAS Institute https://www.sas.com/storefront/aux/en/splsb/65423_excerpt.pdf 2) Williams, Graham, Data Mining with Rattle and R, Springer, 2011 3) my lecture notes
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* *An Introduction to R* http://canisius.edu/~yany/doc/R-intro.pdf *The R Language Definition* http://canisius.edu/~yany/doc/R-lang.pdf My websites: http://www3.canisius.edu/~yany/R.shtml http://www3.canisius.edu/~yany/sas.shtml
QR codes	   
One-line R codes	*> source("http://canisius.edu/~yany/da521.R")* Note: I will explain this line in week 2

[1] I could help you to debug your programs written in R, SAS, Python, Matlab or Perl.

Course Description:	This course focuses on hands-on and term project. It serves as a link between many core courses, such as Data Cleaning, Machines Learning and domain knowledge, such as Economics, Accounting, Finance, and Marketing. Students would apply what they have learnt, such as machine learning, to various real world situations. For students with accounting background, they learn how to process 10-K (annual reports downloaded from SEC's website). For students in Economics, they learn how to generate R and SAS data sets from the data downloaded from the Federal Reserve Bank's Data Library and US Census and apply them to predict the market future trend. For students with a finance background, they learn how to process CRSP and Compustat to evaluate various trading strategies, such as momentum strategy, industry momentum strategy, and 52-high trading strategy. In addition, they learn how to generate various SAS and R data sets from Prof. French's Data Library. For students with marketing knowledge, they learn how to parse social media data to fine tune their marketing strategies. For students from other areas, they learn how to estimate the gender and age group by analyzing million cell phone's usages such as brand, event, timestamp of the events, app downloaded. This course uses two languages: R and SAS. Students are expected to finish a big term project.
Four objectives:	1) review some basic techniques, such as linear regression, generalized linear regression, non-linear regression, logit model, discriminant analysis, principal component analysis, decision tree, random forest, boost, neural network [note: this course would not teach those concepts nor methodologies Instead, I would show how to apply those mythologies to real world situation, i.e., use real-world data] 2) continue to learn SAS and R 3) hands-on and use real-world data 4) finish a big term project
R Software:	R is open source statistical and computational software, see www.r-project.org..
SAS	SAS is powerful tool and it is adopted by many companies and financial institution, especial banks. However, its annual fee is very high. [2]
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.
Course Level	Apply both R and SAS to various real-world situations. In other words,

[2] http://www.sas.com/en_us/home.html

Learning Goals:	students are expected to combine the data analytical skills they have learnt with their domain knowledge. The objective is to finish a big meaningful project by the end of this course.
College, Program and Major Learning Goals:	This course is designed to help students achieve one or more College Core, Business Program and/or Major level learning goals and objectives. You can see the specific College, Program or Major level learning goals and objectives associated with the course from this page on the College website: http://bit.ly/bcoreLG
Grade Evaluation:	Data Cases (6 to 7) 35% Group project 35% Group presentation 15% Class participation 15% ---------- Total 100%
Teaching Methods:	Each class will be consist of two parts: lecture (including discussion of homework) and hands-on.
Group project	Each group can have up to three members. A topic should be closely associated with this course. The maximum number of pages of your report is 15 with 12-point font. Please discuss with me your topic before you start to work on it. Three parts are essential: 1) theory and background of the topic, 2) SAS or R programs with a short explanation of the codes, 3) final data set (plus the codes to process the data, the source of raw data) Note: please do not send me your raw data. Note: see a list of potential topics for the group projects.
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.
Canisius College Academic Calendar	http://www.canisius.edu/academics/events/

Term Project: Each group can have up to three members. A topic should be closely associated with this course. The maximum number of pages of your report is 15 with 12-point font. Please discuss with me your topic before you start to work on it. Some basic criterions are listed below. Real world topics are especially encouraged. Three parts are essential:

1) Theory and background of the topic,
2) SAS or R programs with a short explanation of the codes,
3) Final data set (plus the codes to process the data, the source of raw data) Note: please do not send me your raw data.

Tentative schedule

Week	Date	Topics	Data case
1		Syllabus discussion and Reviews A short survey, self-intro, syllabus, course structure, group project Review of basic concepts (metrologies) Review of R and SAS	
2		Linear regression, T-test, F-test, Test of equal-variance, test of equal means Fama-French-Carhar 4 factor models, Test of January-effect using both R and SAS	#1
3		Introduction to R rattle package Decision tree, logistic regression Small samples for R data set, e.g., `> load(url("http://canisius.edu/~yany/RData/titanic.rda"))`	
4		Clustering analysis, Principal Component Analysis Random tree models using R Test of momentum trading strategy (SAS and CRSP)	#2
5		Boost model using R Which party manages economy better? (CRSP, R or SAS)	
6		Discuss Term Project #1: Momentum trading strategy (SAS and CRSP), see Appendix D.	#3
7		Speed issue More on macro Discuss Term Project #2: Accept or deny a load application	#4
8		SAS merge different data sets Discuss Term project #3: What is a user's next month purchase?	#5
9		Speed issue More on macro Discuss Term project #4: Process TAQ High-frequency data efficiently	#6
10		SAS text mining Discuss Term project #5: How to parse 10-K data? (SEC filings, R and SAS)	#7
11		SAS Marco Discuss Term project #6: Sentiment analysis	
12		SAS discriminate Analysis and random forest and boost Discuss Term project #7: What is a cell phone's user gender and age group based on his/her cell phone's usages? (SAS and TalkingChina data sets, see Appendix E)	
13		SAS merge different data sets Discuss Term project #8: 52-week high-trading strategy	
14		Group presentation	
15		Backup day	

References

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

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

FosterEmail, Kenneth R, Robert Koprowski and Joseph D Skufca, Machine learning, medical diagnosis, and biomedical engineering research – commentary, https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-94

Jegadeesh Narasimhan and Sheridan Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *Journal of Finance* 48 (1), 65-91.

Keeso, Alan, Big Data and Environmental Sustainability: A Conversation Starter, University of Oxford, working paper http://www.smithschool.ox.ac.uk/library/working-papers/workingpaper%2014-04.pdf

Lee, Charles M.C., 1992. Earnings news and small traders: An intraday analysis. Journal of Accounting and Economics 15, 265-302.

Li, Feng, 2008, Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* 45, 221–247.

Loughran, Tim, Bill McDonald, 2015, Textual Analysis in Accounting and Finance: A survey, working paper, University of Notre Dame.

Loughran, Tim, Bill McDonald, 2014, Measuring Readability in Financial Disclosures, *Journal of Finance* 69,4,1643-1671.

Loughran, Tim, Bill McDonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance* 66, 1, 67-97.

Ryan Madden, Ryan, Machine Learning for Predicting The Dow Jones Industrial Average, Northwestern University, working paper, http://ryanmadden.net/ml-dowjones/resources/Madden_FinalReport.pdf

SAS Institute, Combining Knowledge and Data Mining to Understand Sentiment – A Practical Assessment of Approaches.

U.S. Department of Health and Human Services, Big Data to Knowledge (BD2K), https://datascience.nih.gov/bd2k

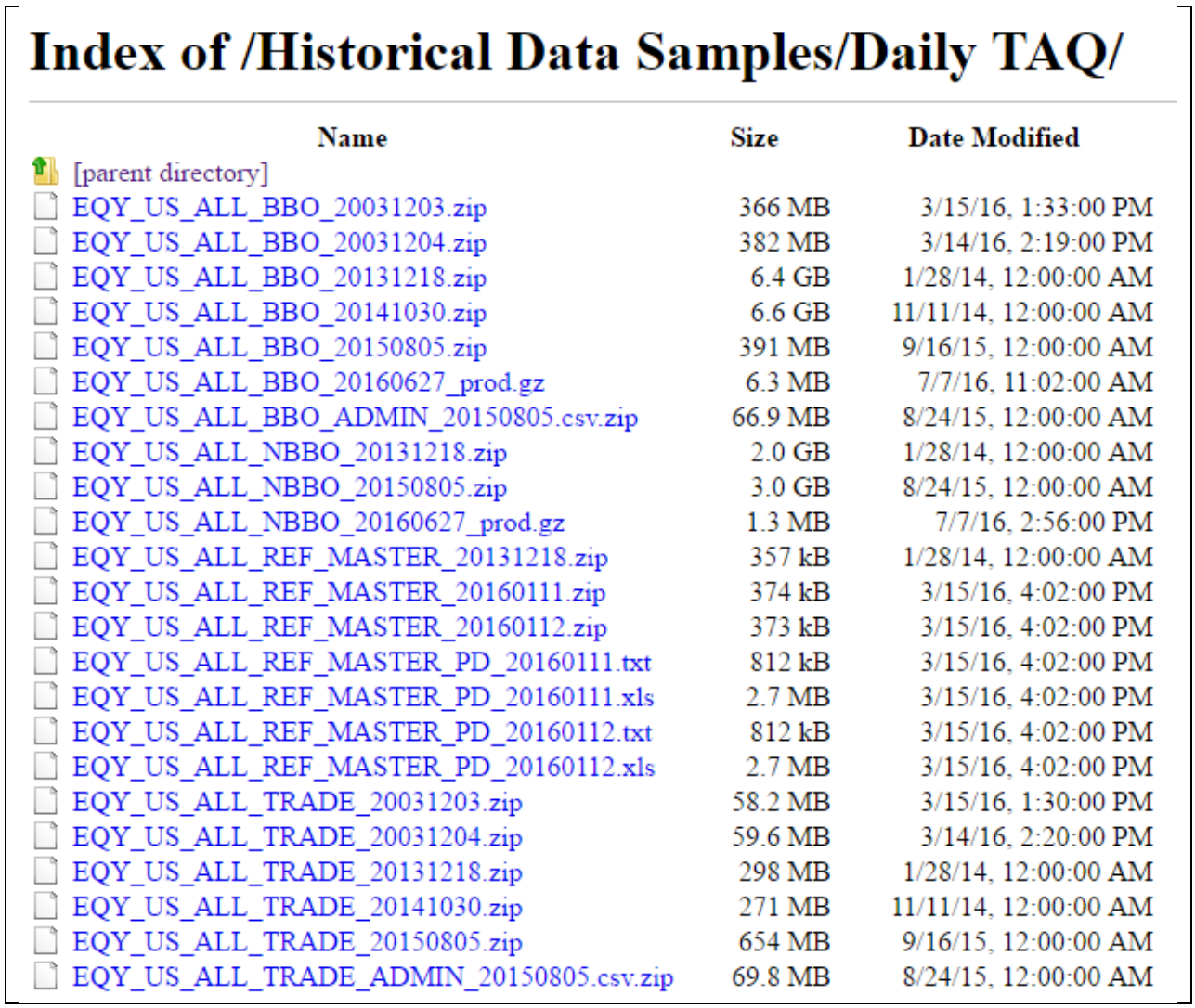
USEA (United States Environmental Protection Agency), Integrating Data from Multidisciplinary Research, Session I: Introducing the Big Picture, https://clu-in.org/conf/tio/IntegratingData1_062415/

Appendix A: Sources of data

Type	Data sources
Open data	SEC filings[3] Note that I have downloaded all 10-K filings (annual reports) from 1993Q1 to 2016Q2. The total size is about 440 G with 210,842 files.
	Prof. French Data Library[4]
	Federal Reserve Banks' data library[5]
	BLS (Bureau of Labor Statistics) [6]
	TDAQ (Daily Trade and Quote) data set (NYSE)
	US Census Bureau[7]
	Healthdata.gov[8]
	Amazon public data sets [9]
	DTAQ data sets (New York Stock Exchange)[10]
	The data sets from Kaggle.com[11]
	UC Irvine Machine Learning Repository [12]
	College data [13]
	Data Science Central[14]
	Opinion mining and sentiment analysis data sets by Prof. Bing Liu[15]
	Text Analysis for finance by Prof. Bill McDonald[16]
	Ecological and Spatial Data Sources[17]
Paid databases	CRSP: trading data for all stocks listed in the US from 1926 onward[18]
	Compustat: accounting information from 1950s onward.[19]
	TAQ (Trade and Quote, high-frequency data). [20] The size of the data set is huge, e.g., the zipped data for 2015/1218 is 2G, see Appendix B.
Competition	Kaggle.com[21]
	Analytics Vidha, Datahack [22]
	Analytics, Data Science, Data Mining Competitions[23]
	Tera Data competition[24]
	Data Science Central competitions[25]

[3] https://www.sec.gov/edgar.shtml
[4] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[5] http://www.federalreserve.gov/econresdata/default.htm
[6] http://www.bls.gov/
[7] http://www.census.gov/data.html
[8] http://www.healthdata.gov/
[9] https://aws.amazon.com/datasets/
[10] ftp://ftp.nyxdata.com/Historical%20Data%20Samples/
[11] https://www.kaggle.com/
[12] https://archive.ics.uci.edu/ml/datasets.html
[13] http://www.collegedata.com/
[14] http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets
[15] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[16] http://www3.nd.edu/~mcdonald/Word_Lists.html
[17] https://www.nceas.ucsb.edu/scicomp/data
[18] http://crsp.com/
[19] http://www.spcapitaliq.com/our-capabilities/our-capabilities.html?product=compustat-research-insight
[20] http://www.nyxdata.com/Data-Products/NYSE-Trades-EOD
[21] https://www.kaggle.com/
[22] http://datahack.analyticsvidhya.com/contest/all
[23] http://www.kdnuggets.com/competitions/
[24] http://www.teradatauniversitynetwork.com/PARTNERS2016/2016-TUN-Data-Challenge/

Appendix B: 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/

[25] http://www.datasciencecentral.com/group/resources/forum/topics/best-kept-secret-about-data-science-competitions

Appendix C: A list of potential topics for term projects

Panel A: General	
1	What is a cell phone user's gender and age group? The total size of 8 csv file is 1.27G. (see Appendix E for more detail)
2	Predict whether income exceeds $50K/year based on census data. Also known as "Census Income" dataset.
3	As a bank clerk, should you grant the loan application or not?
4	Based on their current months' purchases for those 10 produces and other data items, what are their next month's purchases"
5	Credit approval
Panel B: Accounting	
6	Parse all 10-K filings from SEC [note: 10-K is the annual financial statements all public company in US filed with SEC. The total size of 10-K fillings from Q1 1993 to Q2 2016] is 440 G (gigabyte, $1G=10^9$ bite) with 210,842 files.
7	Parse all 13-f filings from SEC [note: 13-f is the quarterly filings by financial institutions reporting their shareholding. The size of 13-K fillings from Q1 1994 to Q2 2016 is 9.39G with 179,043 filings.
8	Readability of 10-K filings the size of all 10-K from Q1 1993 to Q2 2016 is 440G and the number of 10-K fillings is 210,842.
9	The impact of Readability on the quality of a firm
10	Readability and post-earning announcement drift
Panel C: Economics	
11	Generate SAS data sets from US Census [2010 Census Demographic Profile Summary file][26]
12	Sentiment analysis: Which stocks perform better?
13	Which party, Democratic and Republican, could manage economy better?[27]
14	How to design a better market business cycle indicator?
15	Real time GDP estimate based on big data

[26] https://www.census.gov/prod/cen2010/doc/dpsf.pdf
[27] http://www.enchantedlearning.com/history/us/pres/list.shtml

Appendix C (continued)

Panel D: Finance	
16	Is the so-called momentum trading strategy is profitable? [Note: use all stocks from CRSP from 1926 to 2015, the number of stocks is 31,219. The total size of several useful monthly SAS data sets (in SAS format) is 1.0G]
17	Trading direction, Lee and Ready (1991) [Note: the size of one day, 12/18/2013, is 2.3G] [note: the total size of December 2013, 42 zipped file is 29.8G]
18	Process TAQ (NYSE Trade and Quote high-frequency) database How to process DTAQ (NYSE millisecond trading data) Note: the size of zipped one day data is about 2G.
19	Machine Learning based ZZAlpha Ltd. Stock Recommendations 2012-2014 Data Set
20	Machine Learning for Predicting The Dow Jones Industrial Average[28]
21	Insurance Company Benchmark (COIL 2000) [29] Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers.
Panel E: Marketing	
22	How to parse media data? (mining social media data)
23	Bank Marketing Data Set[30]
24	Sentiment analysis and its impact on a firm's market strategy [SAS: Combining Knowledge and Data Mining to Understand Sentiment – A Practical Assessment of Approaches][31]
25	How to design a marketing strategy based on instant feedback?
26	How to design a personalized text message for each potential consumer?
27	Graphical presentation of inter connections (correlation)
28	optimizing supply chain distribution
29	segmenting consumer markets
30	customer lifetime value
31	search engine optimization
32	sales force management optimization
33	Direct marketing and application of uplift models[32]
34	Market segmentation analysis (clustering analysis)

[28] http://ryanmadden.net/ml-dowjones/resources/Madden_FinalReport.pdf
[29] https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29
[30] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
[31] http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/combining-knowledge-data-mining-to-understand-sentiment-105008.pdf
[32] http://www.analyticbridge.com/profiles/blogs/what-are-uplift-models

Appendix C (continued)

Panel F: Healthcare	
35	BRFSS Prevalence And Trends Data: Tobacco Use - Adults Who Are Current Smokers for 1995-2010
36	Selected Trend Table from Health, United States, 2011. Diabetes prevalence and glycemic control among adults 20 years of age and over, by sex, age, and race and Hispanic origin: United States, selected years 1988–1994 through 2003–2006
37	Health Professions Education Foundation Awardees by Zip Code
38	Healthy People 2020 - Leading Health Indicators
39	Census Data - Languages spoken in Chicago, 2008 – 2012
Panel G: Environment	
40	Environmental data mining and modeling based on machine learning algorithms and geostatistics[33]
41	Machine Learning Methods Without Tears: A Primer for Ecologists [34]
Panel H: Life Sciences	
42	Parkinson's Tele-monitoring[35]
43	Heart Disease[36]
Panel H: Social Sciences	
44	Census income (KDD) [37]
45	Communities and Crime Unnormalized [38] (Regression)
46	Blog Feedback (regression) [39]
Panel J: Sports medicine/athletics	
47	Machine learning and sport betting [40]
49	Machine learning for Personized medicine[41]
50	Machine Learning in Health Care [42]

[33] http://ac.els-cdn.com/S1364815203002032/1-s2.0-S1364815203002032-main.pdf?_tid=14da2a1c-5efc-11e6-b51f-00000aab0f27&acdnat=1470834925_e6ad4a66f6eddff5b4e3f494eb6c2ccf
[34] http://www.jstor.org/stable/pdf/10.1086/587826.pdf?_=1470834864323
[35] https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring
[36] https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[37] https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
[38] https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized
[39] https://archive.ics.uci.edu/ml/datasets/BlogFeedback
[40] http://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf
[41] http://www.mlpm.eu/
[42] http://mucmd.org/

Appendix D: One example of potential term project #2

Momentum strategy[43]

Objectives: 1) Understand CRSP database

2) Understand how to use SAS, Matlab or R to process CRSP data

3) Prove or disapprove so-called momentum strategy by replicating Table 1 of Jegadeesh and Titman (1993)

Size of data: use all stocks from CRSP from 1926 to 2015, the number of stocks is 31,219. The total size of several monthly SAS data sets, in SAS format, is about 1.0G.

Basic logic: According to Jegadeesh and Titman (1993) it is a profitable trading strategy if we buy the past winners and sell the past losers.

Notations: Check the past K-month returns, and then form a portfolio for L months, Where K=3,6,9 and 12 and L=3, 6, 9 and 12. Below we use K=L=6 as an example.

Trading strategy: Estimate all stock's past 6- month returns and sort stocks into 10 groups (deciles) according to them. Long the top decile (winners) and short the bottom decile (losers) for the next 6 months.

Procedure:

Step 0: Starting month: January 1965

Step 1: Retrieve CRSP data (PERMNO, DATE and RET)

Step 2: Estimate past 6-month cumulative returns R_t^{6month}

Step 3: Sort all stocks into deciles according to this cumulative 6-month returns

Step 4: Long winners (best return group) and short losers for the next 6-month

Step 5: Estimate portfolio returns

Step 6: Move to the next month and repeat the above steps until the last month

References

Jegadeesh Narasimhan and Sheridan Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *Journal of Finance* 48 (1), 65-91.

[43] 4/22/2013, yuxing.yan@hofstra.edu

Table 1 from Jegadeesh and Titman (1993).

Table I

Returns of Relative Strength Portfolios

The relative strength portfolios are formed based on J-month lagged returns and held for K months. The values of J and K for the different strategies are indicated in the first column and row, respectively. The stocks are ranked in ascending order on the basis of J-month lagged returns and an equally weighted portfolio of stocks in the lowest past return decile is the *sell* portfolio and an equally weighted portfolio of the stocks in the highest return decile is the *buy* portfolio. The average monthly returns of these portfolios are presented in this table. The relative strength portfolios in Panel A are formed immediately after the lagged returns are measured for the purpose of portfolio formation. The relative strength portfolios in Panel B are formed 1 week after the lagged returns used for forming these portfolios are measured. The t-statistics are reported in parentheses. The sample period is January 1965 to December 1989.

		Panel A				Panel B			
J		K = 3	6	9	12	K = 3	6	9	12
3	Sell	0.0108	0.0091	0.0092	0.0087	0.0083	0.0079	0.0084	0.0083
		(2.16)	(1.87)	(1.92)	(1.87)	(1.67)	(1.64)	(1.77)	(1.79)
3	Buy	0.0140	0.0149	0.0152	.0156	0.0156	0.0158	0.0158	0.0160
		(3.57)	(3.78)	(3.83)	(3.89)	(3.95)	(3.98)	(3.96)	(3.98)
3	Buy-sell	0.0032	0.0058	0.0061	0.0069	0.0073	0.0078	0.0074	0.0077
		(1.10)	(2.29)	(2.69)	(3.53)	(2.61)	(3.16)	(3.36)	(4.00)
6	Sell	0.0087	0.0079	0.0072	0.0080	0.0066	0.0068	0.0067	0.0076
		(1.67)	(1.56)	(1.48)	(1.66)	(1.28)	(1.35)	(1.38)	(1.58)
6	Buy	0.0171	0.0174	0.0174	0.0166	0.0179	0.0178	0.0175	0.0166
		(4.28)	(4.33)	(4.31)	(4.13)	(4.47)	(4.41)	(4.32)	(4.13)
6	Buy-sell	0.0084	0.0095	0.0102	0.0086	0.0114	0.0110	0.0108	0.0090
		(2.44)	(3.07)	(3.76)	(3.36)	(3.37)	(3.61)	(4.01)	(3.54)
9	Sell	0.0077	0.0065	0.0071	0.0082	0.0058	0.0058	0.0066	0.0078
		(1.47)	(1.29)	(1.43)	(1.66)	(1.13)	(1.15)	(1.34)	(1.59)
9	Buy	0.0186	0.0186	0.0176	0.0164	0.0193	0.0188	0.0176	0.0164
		(4.56)	(4.53)	(4.30)	(4.03)	(4.72)	(4.56)	(4.30)	(4.04)
9	Buy-sell	0.0109	0.0121	0.0105	0.0082	0.0135	0.0130	0.0109	0.0085
		(3.03)	(3.78)	(3.47)	(2.89)	(3.85)	(4.09)	(3.67)	(3.04)
12	Sell	0.0060	0.0065	0.0075	0.0087	0.0048	0.0058	0.0070	0.0085
		(1.17)	(1.29)	(1.48)	(1.74)	(0.93)	(1.15)	(1.40)	(1.71)
12	Buy	0.0192	0.0179	0.0168	0.0155	0.0196	0.0179	0.0167	0.0154
		(4.63)	(4.36)	(4.10)	(3.81)	(4.73)	(4.36)	(4.09)	(3.79)
12	Buy-sell	0.0131	0.0114	0.0093	0.0068	0.0149	0.0121	0.0096	0.0069
		(3.74)	(3.40)	(2.95)	(2.25)	(4.28)	(3.65)	(3.09)	(2.31)

Appendix E: Another example of potential term project (#10)

The objective of this project is to identify a cell phone's user's gender and age group. The total size of 8 csv file is 1.27G.

File Name	Available Formats
gender_age_test.csv	.zip (1.05 mb)
app_labels.csv	.zip (4.04 mb)
label_categories.csv	.zip (7.67 kb)
phone_brand_device_model.csv	.zip (2.42 mb)
sample_submission.csv	.zip (1.32 mb)
events.csv	.zip (62.24 mb)
app_events.csv	.zip (211.27 mb)
gender_age_train.csv	.zip (891.47 kb)

```
-rw-r--r-- 1 abced xyx 1037267659 Jul 29 16:45 app_events.csv
-rw-r--r-- 1 abced xyx   11190003 Jul 29 16:45 app_labels.csv
-rw-r--r-- 1 abced xyx  195433779 Jul 29 16:45 events.csv
-rw-r--r-- 1 abced xyx    2284333 Jul 29 16:46 gender_age_test.csv
-rw-r--r-- 1 abced xyx    2366486 Jul 29 16:49 gender_age_train.csv
-rw-r--r-- 1 abced xyx      16450 Jul 29 16:46 label_categories.csv
-rw-r--r-- 1 abced xyx    6715635 Jul 29 16:46
phone_brand_device_model.csv
-rw-r--r-- 1 abced xyx   11698373 Jul 29 16:47 sample_submission.csv
```

A few lines from teach data sets are given below.

app_events.csv

```
event_id,app_id,is_installed,is_active
2,5927333115845830913,1,1
2,-5720078949152207372,1,0
2,-1633887856876571208,1,0
2,-653184325010919369,1,1
2,8693964245073640147,1,1
2,4775896950989639373,1,1
2,-8022267440849930066,1,0
2,9112463267739110219,1,0
2,-3725672010020973973,1,0
```

app_labels.csv

```
app_id,label_id
7324884708820027918,251
-4494216993218550286,251
6058196446775239644,406
6058196446775239644,407
8694625920731541625,406
8694625920731541625,407
1977658975649789753,406
1977658975649789753,407
7311663864768030840,256
```

gender_age_train.csv

```
device_id,gender,age,group
-8076087639492063270,M,35,M32-38
-2897161552818060146,M,35,M32-38
-8260683887967679142,M,35,M32-38
-4938849341048082022,M,30,M29-31
245133531816851882,M,30,M29-31
-1297074871525174196,F,24,F24-26
236877999787307864,M,36,M32-38
-8098239495777311881,M,38,M32-38
176515041953473526,M,33,M32-38
```

gender_age_test.csv

```
device_id
1002079943728939269
-1547860181818787117
7374582448058474277
-6220210354783429585
-5893464122623104785
-7560708697029818408
289797889702373958
-402874006399730161
5751283639860028129
```

label_categories.csv

```
label_id,category
1,
2,game-game type
3,game-Game themes
4,game-Art Style
```

```
5,game-Leisure time
6,game-Cutting things
7,game-Finding fault
8,game-stress reliever
9,game-pet
```

phone_brand_device_model.csv

```
device_id,phone_brand,device_model
-8890648629457979026,-9,-9
1277779817574759137,-9,MI 2
5137427614288105724,-9,Galaxy S4
3669464369358936369,SUGAR,-9
-5019277647504317457,-9,Galaxy Note 2
3238009352149731868,-9,Mate
-3883532755183027260,-9,MI 2S
```

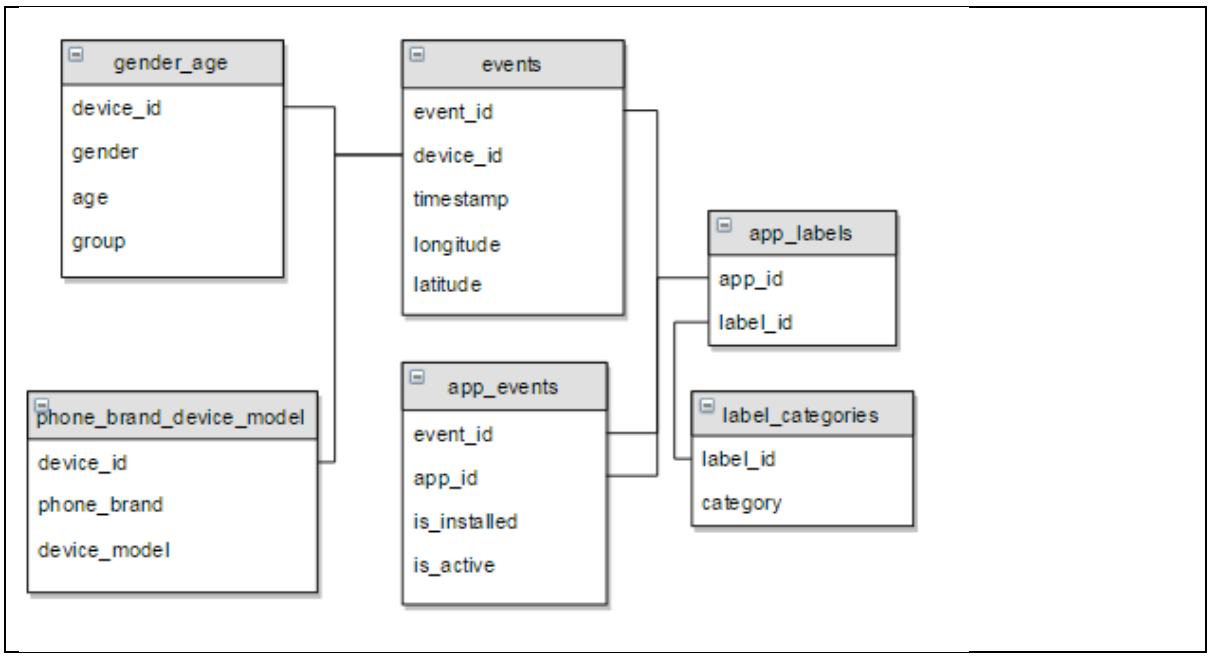
events.csv

```
event_id,device_id,timestamp,longitude,latitude
1,29182687948017175,2016-05-01 00:55:25,121.38,31.24
2,-6401643145415154744,2016-05-01 00:54:12,103.65,30.97
3,-4833982096941402721,2016-05-01 00:08:05,106.60,29.70
4,-6815121365017318426,2016-05-01 00:06:40,104.27,23.28
5,-5372797595892518570,2016-05-01 00:07:18,115.88,28.66
6,1476664663289716375,2016-05-01 00:27:21,0.00,0.00
7,5990801147117726237,2016-05-01 00:15:13,113.73,23.00
8,1782450055857303792,2016-05-01 00:15:35,113.94,34.70
9,-2073340001552902943,2016-05-01 00:15:33,0.00,0.00
```

sample_submission.csv

```
device_id,F23-,F24-26,F27-28,F29-32,F33-42,F43+,M22-,M23-26,M27-28,M29-31,M32-38,M39+
1002079943728939269,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
-1547860181818787117,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
7374582448058474277,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
-6220210354783429585,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
-5893464122623104785,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
-7560708697029818408,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
289797889702373958,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
-4028740063997301 61,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
5751283639860028129,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833
```

The data sets are interconnected in the following graph.



Appendix F: Example of potential term project (#8) Estimate the readability of 10-K filings

Objectives:

1) Understand the usage of 10-K

2) learn how to parse 10-K

3) understand the Fog-index and learn how to calculate it for each 10-K filing

4) Comments on your result

Source of data

a) SEC EDGAR (Electronic Data Gathering , Analysis and Retrieval)

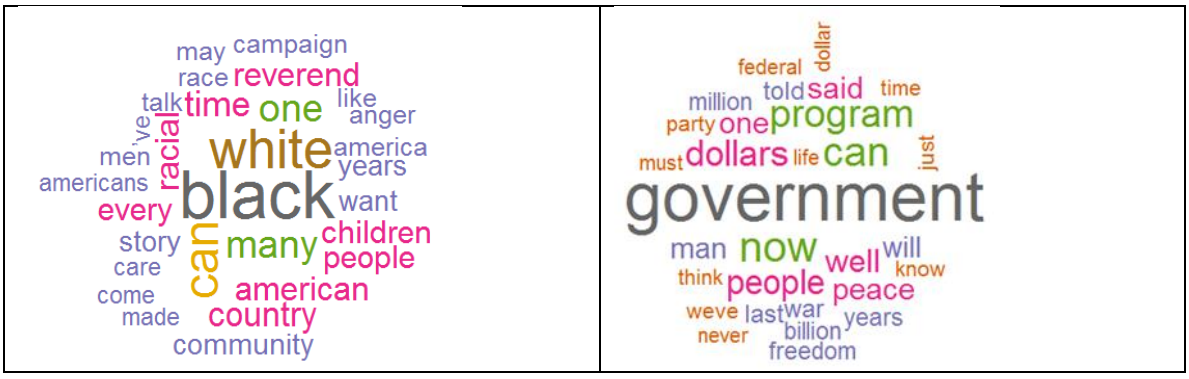
b) I have all 10-K filings from Q1 1993 to Q2 2016 (the number of filings is 210,842 and the size is 440G)

Structure vs. unstructured data

The unstructured information has a lion share of all information, 70% to 80% and it is reported that 80% of structured information came from unstructured one. On the other-hand, SEC filings is an important source of information (gold mine) since public companies, certain insiders, and broker-dealers are required to make regular SEC filings.

Text analysis

Text is one of the most important informant belongs to unstructured information. Text analysis, also called text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. For example, we could look at the frequency of each words, keywords, number of lines, sentences, frequency of positive words vs. negative, tone of the speech etc. For example, let's look at the top used words by Reagan in 1994 and Obama 2008, see below. Which one belongs to Obama?



Text analysis in accounting

Li (2008) shows that the readability, defined by Fog Index, of 10-K filings has a statistically significant impact on the performance of a firm's subsequent years.

$$Fog\ index = 0.4 * (n + p) \quad (1)$$

where n is the average number of words per sentence, while p is the percentage of complex words compared with all words. A complex word is a word that has more than two syllables. The following Appendices A and B are from Li (2008).

Appendix A. Steps to calculate the readability indices

This appendix explains the details of calculating the readability indices starting from the raw 10-K filings used in this paper. I first download the 10-K report from Edgar and perform the following editing before further analysis. First, the heading information that is contained between ⟨SEC-HEADER⟩ and ⟨/SEC-HEADER⟩ is deleted. Second, all the tables that begin with ⟨TABLE⟩ and end with ⟨/TABLE⟩ or the paragraphs that contain ⟨S⟩ or ⟨C⟩ are deleted, because ⟨S⟩ and ⟨C⟩ tags are used by some firms to present tables. Next, all the tags in the format of ⟨…⟩ and ⟨&…⟩, which are used widely in documents in SEC HTML or XML format documents, are replaced with blanks. Finally, to make sure that all the tables, tabulated text, or financial statements are excluded, all the paragraphs with more than 50% of non-alphabetic characters (e.g., white spaces or numbers) are deleted.

The file after the editing is then analyzed using the Fathom package in Perl. The package can calculate the typical text statistics, including the number of characters, number of words, percent of complex words (i.e., words with more than three syllables), number of sentences, number of text lines, number of paragraphs, syllables per word, and words per sentence. Based on the statistics, the package also produces the summary readability indices used in the paper.

Appendix B. Steps to extract MD&A and Notes to the financial statements

This appendix explains the details of extracting the MD&A section and Notes from 10-K filings. Starting with the raw 10-K file, I first delete the SEC-header information, all the contents between ⟨TABLE⟩ and ⟨/TABLE⟩ text, the paragraphs that contain ⟨S⟩ or ⟨C⟩, all the tags in the format of ⟨…⟩ and ⟨&…⟩ are removed using the same process described in Appendix A.

Within the remaining text, the program identifies a line that satisfies one of the following criteria as the *beginning* of the MD&A section: (1) the line starts with "management's discussion" or "management's discussion" following some white spaces; (2) the line contains "management's discussion" and ("item" + one or more white space + "7") and does not contain the word "see"; (3) the line starts with some white spaces followed by "managements discussion" or "managements discussion"; or (4) the line contains "managements discussion" and ("item" + one or more white space + "7") and does not contain the word "see." Since many firms refer to the MD&A section in the front-matter of the annual reports, the word "see" serves to identify all such situations. The program identifies a line that satisfies one of the following criteria as the *ending* of the MD&A section: (1) the line begins with some white spaces followed by "Financial Statements" or "Financial Statements"; (2) the line contains "item" followed by one or more white spaces and the number "8"; (3) the line contains "Supplementary Data"; or (4) the line begins with some white spaces followed by "SUMMARY OF SELECTED FINANCIAL DATA" or "SUMMARY OF SELECTED FINANCIAL DATA." Most firms have a table of contents listing the main sections of the 10-K filing. In some instances, this table of contents is not embedded between ⟨TABLE⟩ and ⟨/TABLE⟩ and therefore is not cleaned in the previous steps. As a result, the line in the table of contents about MD&A will also be picked up by the program as part of the MD&A.

Similarly, the program identifies a line as the *beginning* of the Notes, if: (1) the line starts with "NOTES TO" or some white spaces followed by "NOTES TO"; and (2) the line does not contain any number except when it follows "for the years ended." The program identifies a line that satisfies one of the following criteria as the *ending* of the Notes: (1) the line contains "Changes in and Disagreements with Accountants" or "DISAGREEMENTS ON ACCOUNTING"; (2) the line contains "DIRECTORS AND EXECUTIVE OFFICERS"; or (3) the line contains "exhibit index."

After the MD&A and the Notes are identified, all the paragraphs with more than 50% of non-alphabetic characters (e.g., white spaces or numbers) are deleted. Finally, the Fathom package is used to calculate the readability measures.

Reference

Li, Feng, 2008, Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* 45, 221–247.

Appendix G: Example of potential term project (#18) Trading direction (Lee and Ready, 1991)

Objectives:

1) understand the structure of TAQ
2) understand the Lee-Ready (1991) methodology to identify trading direction
3) understand how to process big data set
4) using SAS to implement the methodology by using one month's data (30G) [44]

TAQ stands for Trade and Quote database and it is developed and maintained by New York Stock Exchange. TAQ is a a high-frequency trading database. It has second-by-second CT (Consolidated Trading) and CQ (Consolidated Quote) data. Below are examples for CT and CQ. Size is a big issue for processing TAQ database: one month's TAQ data is about 40G.

Table 1: The first several lines of CT are shown below.

Obs	SYMBOL	DATE	TIME	PRICE	SIZE	G127	CORR	COND	EX	TSEQ
1	A	20000103	9:34:01	78.75	64700	40	0		N	807127
2	A	20000103	9:34:04	78.75	100	0	0		M	0
3	A	20000103	9:34:04	78.75	1000	0	0		M	0
4	A	20000103	9:34:04	78.75	100	0	0		M	0
5	A	20000103	9:34:04	78.75	200	0	0		M	0
6	A	20000103	9:34:04	78.75	100	0	0		M	0
7	A	20000103	9:34:04	78.75	100	0	0		M	0
8	A	20000103	9:34:04	78.75	100	0	0		M	0
9	A	20000103	9:34:04	78.75	100	0	0		M	0
10	A	20000103	9:34:04	78.75	100	0	0		M	0

Definitions of variables from CT data set:

SYMBOL this variable is not a permanent stock

G127 Combination of following 3 rules (G rule: trading for its own account, 127 rule: executed as a block position. Stopped stock indicator

e.g., G127=0, does not qualify as "G", Rule 12 or stopped stock trade

G127=40 A display book-reported trade

CORR Correction indicator, e.g., CORR=0 regular trade

COND Condition of a trade e.g., COND='A' Cash-only basis

The first several lines of CQ are given below.

Obs	SYMBOL	DATE	TIME	BID	OFR	BIDSIZ	OFRSIZ	MODE	EX	MMID	SEQNUM
1	A	20000103	8:59:07	0.000	0.000	0	0	12	T	PTRS	0
2	A	20000103	8:59:07	0.000	0.000	0	0	12	T	SWST	0
3	A	20000103	8:59:07	0.000	0.000	0	0	12	T	TRIM	0
4	A	20000103	8:59:07	0.000	0.000	0	0	12	T	MADF	0
5	A	20000103	9:34:02	0.000	0.000	0	0	12	C		0
6	A	20000103	9:34:08	78.625	78.875	10	10	10	N		807129
7	A	20000103	9:34:10	78.500	79.000	1	1	12	X		0
8	A	20000103	9:34:10	77.750	79.750	1	1	12	C		0
9	A	20000103	9:34:12	78.500	79.000	1	1	12	T	MADF	0
10	A	20000103	9:34:12	78.500	79.000	1	1	12	T	CAES	0

[44] Just image to estimate all trading direction from 1993 to 2014.

Definitions for CQ data set:

BID	Bid price
OFR	Offer price
BIDSIZ	Bid size (100 share units)
OFRSIZ	Offer size (100 share units)
MODE	Quote condition
	e.g. MODE=0 Invalid field
	MODE=4 News dissemination (regulatory halt)
MMID	NASDAQ market maker

However, there is no indicator whether a trade is buyer-initiated or seller initiated. Lee and Ready (1991) develop a methodology to identity who initiated a trade. Their methodology has two parts: Quote test first. If it fails then a Tick test. Identify who initiates a trade

Quote test: Seller initiated: price < (bid + ask)/2

Buyer-initiated: price > (bid + ask)/2

Note: that when prices =(bid+ask)/2 then use tick test

Tick test: Seller-initiated: P(t) <P(t-1)

Buyer-initiated : p(t)>P(t-1)

Below is one procedure:

Step 1: Filtering out invalid trades

Keep if 1) price: price >0
2) size : size >0
3) CORR: Correction indicator CORR = 0, 1 or 2
4) COND : Sale Condition
COND not in ("O" "Z" "B" "T" "L" "G" "W" "J" "K")

Step 2: Filtering out invalid quotes

Keep if 1) price: bid >0, ofr >0
2) size : bidsiz>0 , ofrsiz>0
3) mode: mode not in (4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28)
e.g., mode=4: regulatory halt (news dissemination)
mode=7: non-regulatory halt (order imbalance)
mode=9: regulatory halt

Step 3: Matching trades with quotes (5 second rule)

Step 4: Conduct the Lee and Ready test

Reference

Lee, Charles M.C., 1992. Earnings news and small traders: An intraday analysis. Journal of Accounting and Economics 15, 265-302.