MIS 447/680 – Practical Computing for Data Analytics

(Advanced analytics with R and Python)

Winter - 2017

Mark Isken (isken@oakland.edu) Class meets: Tu 6:30p-9:20p in 327EH
Office: 317 Elliott Hall (248.370.3296)

My OU website: http://www.sba.oakland.edu/faculty/isken/
My hselab website: http://hselab.org/
My LinkedIn Profile: www.linkedin.com/pub/mark-isken/a/633/48/

OVERVIEW

The overlapping fields of business analytics and data science are getting all kinds of attention in the business world. While trusty tools like Excel spreadsheets and SQL databases are still widely used, there are many other tools wielded by analysts. In this class you will begin to learn how to use R and Python (along with a myriad of companion libraries) to do business analytics work such as:

- accessing large datasets and doing the "wrangling" needed to prepare them for analysis
- managing analytical datasets
- exploratory data analysis and visualization
- build predictive models using statistics and machine learning techniques
- create and manage reproducible analytical processes and workflows
- communicate results and tell stories with data and models

Should I take this class? What are the prereqs?

Business analytics work is challenging. Some of you may have taken my MIS 443/546: Business Analytics course. While it takes a spreadsheet based modeling approach, that course illustrates the wide range of skills needed by the modern business analyst, including but not limited to: power spreadsheeting, mathematical modeling, data preparation, statistics, data visualization, VBA programming, database/SQL, and technical communication. Likewise, this course will be challenging but in a really good way.

The size and scope of datasets that business analysts work with now have necessitated an increased focus on the computational aspects of business analytics. That means programming. R is a programmatically based computing environment for doing statistical analysis. Python is a full-fledged programming language that also happens to be widely used in the analytics world. This course will introduce you to this new world of computational analytics. You will pick up valuable and marketable skills and will be well positioned for the continuous learning that is needed to succeed in the analytics field.

This course is open to both undergraduates and graduate students. In terms of prereqs, you should have had an introduction to at least one general purpose programming language. The amount of Excel VBA that I covered in my MIS 546 is a fine level of preparation. In addition, it is important that you’ve had some sort of probability and statistics based quantitative methods course such as business statistics, management science, or my MIS 443/546 course. However, the most important prereqs are interest, determination, willingness to experiment and persist in the face of challenging problems, and a passion to push your analytics skills to a whole new level. This is a special topics course, and as such, I’m assuming you want to be here to actually learn this stuff and not just to satisfy some credit hours requirement. There are much easier courses you could take for checking some box on your degree checklist.

An Upfront Acknowledgement

There are some very good data science courses out there. I’ve been particularly influenced by, and have learned a lot from, the cs109 course at Harvard University. I strongly encourage you to check it out. I will be borrowing numerous ideas and some content from that course including some items from their syllabus. It’s a really good course. Don’t worry, the math and programming requirements for that course are much tougher than ours and I will NOT be trying to duplicate that course here. They also have an army of teaching assistants to run separate lab sessions for teaching programming and can use the lecture sessions for the data science concepts (and for that they have two professors). I have to do it all, and as such, I think our course will be a nice way to prepare for a course like cs109. After our course is finished, I strongly encourage you to try to work through the cs109 homework assignments. They are interesting, challenging, and will provide a natural next step for you in learning how to do analytics work.

TEXTBOOKS

Traditional textbooks don’t exist for a class like this. Instead we’ll be using a number of inexpensive paperback books that will cover different aspects of the course. They are all well worth owning. In addition, we will be using numerous free webbased resources. For example:

- check out my blog post on Learning Python suggestions for business analytics students and professionals

- a great, free book, Introduction to Statistical Learning (with applications in R), that will provide the mathematical background for the algorithms we'll discuss.

Required texts for Winter 2017

All three books have official websites from which you can buy print, PDF, or eBooks. Of course, you can also find them at numerous places on the web. I've listed approximate pricing from checking a few of the online booksellers.

R for Everyone - http://www.jaredlander.com/r-for-everyone/
Jared Lander
~$25 new, less for used

This provides an accessible, modern and thorough introduction to the world of the R statistical computing platform. I've used this book the past two years.

Practical Data Science with R - https://www.manning.com/books/practical-data-science-with-r
Nina Zumel and John Mount
~$40 new, less for used

This is a newish book (2014) that does just what the title suggests. It is structured around typical business analytics or data science projects and covers the main statistical learning techniques along with tons of practical advice on doing data science projects.

Data Wrangling with Python - http://shop.oreilly.com/product/0636920032861.do
Jacqueline Kazil & Katharine Jarmul

~$44 new + ebook, less for used

This is a brand new book (2016) book that I'm really excited about. Finally, a problem driven book that introduces the Python language as it's needed to solve these problems. Tons of practical advice and written in a style that matches how this work is really done - lots of trying stuff and partially succeeding and then trying other stuff ... (repeat till happy). I believe this is a great way to learn to be an effective programmer and both get useful things done and have fun while doing it.

Books I used for the 2014 and 2015 versions of this class

I really like all the books below, but they are **NOT** required for the class. In addition to the books below, we also used R for Everyone.

Doing Data Science: Straight Talk from the Frontline - http://shop.oreilly.com/product/0636920028529.do

Cathy O'Neil & Rachel Schutt
~$25

This book is more a collection of chapters written by the authors and various data science practitioners. It's very readable and full of insights on the practice of data science.

Practical Computing for Biologists - http://practicalcomputing.org/

Steve Haddock & Casey Dunn

~$40-55

I fell in love with this book immediately and found myself wishing that someone would write a similar book for business. It is aimed at scientists who realize that they need to get better at computing to deal with all the data they need to process and analyze. That sounds like many business analysts. It's Mac and Linux based and is crammed full of useful information on text files, using the command line, regular expressions, shell scripts, Python programming, dealing with image files, relational databases and even working with physical data collection devices. Highly, highly recommended.

Python for Data Analysis - http://shop.oreilly.com/product/0636920023784.do

Wes McKinney
~$24

This is a somewhat more advanced book on using Python for data analysis. It was written by the developer of the hugely popular Python package, pandas. In addition to a thorough coverage of pandas, it covers numpy, IPython, and even an intro to the Python language. It's getting a bit dated but is still considered the pandas bible and is a must have for any Pythonista.

SOFTWARE

Business schools tend to be Microsoft dominated places. After all, Excel is the “Swiss army knife” of business and Powerpoint is everywhere. However, the analytics world is a far more diverse place. I’m going to give you an opportunity to explore a wide range of new tools and computing environments. I want all of us to be able to work in the same computing environment whether we are in the lab or at home. So, I’ve created a virtual machine based “analytic appliance” that we’ll call **pcda**. The **pcda** appliance comes preconfigured with:

- A flavor of the Linux OS called Lubuntu
- R and R Studio
- The Anaconda distribution of Python for scientific computing
- Geany, a nice text editor and programming IDE
- PyCharm - a great Python IDE
- A web browser, file manager, command shell, and other common tools

Linux!? Yep, you are going to learn Linux. You may have heard of Ubuntu as it’s the most popular Linux “distro” out there for the average home user. Lubuntu is a lightweight version of Ubuntu that just has the minimal set of the Linux OS that we’ll need. While Lubuntu is GUI based, we will also be using the “shell” (like a Windows command line but a jillion times better and more powerful). Lubuntu (and Ubuntu) are both free and open source.

Both R and Python are free and open source products with huge communities of analytics users and contributors. They overlap to some degree but have distinct strengths. Both are well worth learning. They both allow you to do things that would be absolutely hideous and painful to do in Excel. The **pcda** appliance was created with VirtualBox, a free software package from Oracle for creating and using virtual machines. **pcda** will be available in the 327EH lab and I'll be showing you how you can use it on your own computer as well. As a start, you'll need to download and install VirtualBox.

COURSE STRUCTURE

Lab Sessions

We will meet once per week, Tue from 6:30-9:20 in 327EH. I envision the sessions to be similar to my MIS 443/546 class much more of a "studio" than a lecture class. There will usually be hands-on, interactive lessons where I present various topics and we bring them to life together with software tools. We'll do problem solving, guided tutorials, and discussion of the business analytics and data science worlds. I'll leave a large chunk of time each session for individual and group work during which I'll be the "roving consultant", answering questions and helping you figure things out for yourself. A tentative schedule of topics is near the bottom of this syllabus.

Course Websites

For those of you who've had my MIS 443/546 class, you know that I create extensive Moodle sites with tons of learning resources, files for use in class, the assignments, forums, and whatever else I decide to create. It is pretty likely that I will be creating some screencasts to explain or demonstrate some specific thing that you might want to revisit outside of class. In addition to the course Moodle site, I'll be referencing some items on my hselab.org site as well.

Homework

There will be a number of assignments that will give you a chance to apply the things you've learned in the class and to test your understanding of the material. The goal is learning. The course schedule includes required readings. The goal of the reading assignments is to prepare for class, to familiarize yourself with new terminology and definitions, and to determine which part of the subject needs more attention. The homework assignments may contain questions about the mandatory readings. When answering those please be brief and to the point!

Online Quizzes and In Class Assignments

There will likely be a handful of these kinds of things as I've found them to be a useful incentive for keeping up with the material.

Project

I want to give you the opportunity to apply your new skills to a dataset of interest to you. So, instead of a final exam, you'll have a final project in which you'll identify a question of interest for which relevant data exists. You'll design and create analytical scripts in either R or Python (or both if you wish) to do the analysis, and then create various summaries and visualizations. You'll figure out the best way to communicate the results and findings. For the project, you can form teams of 1-4 students. More information about the final project is available in Moodle.

ASSESSMENT AND GRADING

There are three major components to grading:

- **Homework assignments** 60%
- **In-class assignments and online quizzes** 10%
- **Final project** 30%

I will evaluate your work holistically beyond mechanical correctness and focus on the overall quality of the work. In addition to the scores I will try to give some detailed written feedback. I really don't want to put too much focus on grades. This is a new course, a challenging course, and I really just want people to start to learn this stuff so that they can compete on the analytics job market. So, I'm not going to take a very hard line on grading. I understand that this stuff is hard.

Collaboration Policy

You are welcome to discuss the course's ideas, material, and homework with others in order to better understand it, but the work you turn in must be your own (or for the project, yours and your teammate's). For example, you must write your own code, run your own data analyses, and communicate and explain the results in your own words and with your own visualizations. You may not submit the same or similar work to this course that you have submitted or will submit to another. Nor may you provide or make available solutions to homeworks to individuals who take or may take this course in the future.

Quoting Sources

You must acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header). You can also acknowledge sources in a README.txt file if you used whole classes or libraries. Do not remove any original copyright notices and headers. However, you are encouraged to use libraries, unless explicitly stated otherwise! You may use examples you find on the web as a starting point, provided its license allows you to reuse it. You must quote the source using proper citations (author, year, title, time accessed, URL) both in the source code and in any publicly visible material. You may not use existing complex combinations or large examples. For example, you may not use a ready to use multiple linked view visualization. You may use parts out of such examples.

Missed Activities and Assignment Deadlines

Projects and homework must be turned in on time, with the exception of late days for homework as stated below. It is important that everybody attends and proactively participates in class and online. We understand, however, that certain factors may occasionally interfere with your ability to participate or to hand in work on time.

Homework Deadlines and Late Days

Each student is given six late days for homework at the beginning of the semester. A late day extends the individual homework deadline by 24 hours without penalty. No more than two late days may be used on any one assignment. Assignments handed in more than 48 hours after the original deadline get a max grade of 80%. Late days are intended to give you flexibility: you can use them for any reason –no questions asked. You don't get any bonus points for not using your late days. Also, you can only use late days for the individual homework deadlines – all other deadlines (e.g., project milestones) are hard.

TENTATIVE SCHEDULE - DETAILS SUBJECT TO CHANGE AS WE GO (SEE MOODLE FOR MORE DETAILS)

Topic	Textbook Readings	Date	Week	Day	Subtopic 1	Subtopic 2	Subtopic 3	Subtopic 4
Intro to course and analytical computing, the Linux shell	DWwP App C	**01/10/17**	**1**	**Tue**	Intro to DS and course	pcda appliance	peek at Python, IPython, markdown, shell	Linux shell basics
Intro to Python	DWwP 1, 2, App A, B	**01/17/17**	**2**	**Tue**	Finish shell basics if needed	Python basics	Getting started with IPython notebooks, variables, expressions, print, containers, basic array manipulation in numpy, functions	Python – reading CSV files
Reading text and Excel data, data acquisition	DWwP 3,4,6	**01/24/17**	**3**	**Tue**	Text files	Python – reading files	Python – loops, if-then, list indexing	Python – Reading Excel. range, indexing/slicing, nested loops, counters, break, commenting.
Data cleanup and workflow scripting	DWwP 7, 8	**01/31/17**	**4**	**Tue**	Header replacement using looping and dictionaries	String formatting and datetimes	regex basics	Creating a read, clean, prep, EDA workflow script
EDA - Group by and plotting, workflow scripting, git for version control	DWwP 9, 10	**02/07/17**	**5**	**Tue**	pandas	matplotlib	git and Github	
Data from the web	DWwP 11, 12, 13	**02/14/17**	**6**	**Tue**	web scraping	web APIs		
No Class - Winter Break		**02/21/17**	**7**	**Tue**				
Intro to R and R Studio	RfE 1-6, PDSwR 1-2	**02/28/17**	**8**	**Tue**	Overview of R	vectors, dataframes, getting data into R	Using R Studio	Packages
Summary stats, group by and plotting with R	RfE 7, 8, 11, 12, 15, PDSwR 3, 4	**03/07/17**	**9**	**Tue**	R Studio projects, summary statistics	R plotting, ggplot2	Group by w/ plyr, dplyr	Reshaping
Modeling 1 - Overview and regression	PDSwR 1, 5, 7.1, RfE 16-18	**03/14/17**	**10**	**Tue**	Data science process	Linear regression	model assessment	Using R from Python
Modeling 2 - Basic classifiers	RfE 17.1, PDSwR 6, 7.2	**03/21/17**	**11**	**Tue**	kNN and other “memorization methods”	Logistic regression		
Modeling 3 - Unsupervised methods and scikit-learn	PDSwR 8.1, RfE 22	**03/28/17**	**12**	**Tue**	Cluster analysis	scikit-learn intro (Python)		
Modeling 4 - Trees and forests	PDSwR 9, RfE 20.4, 20.5	**04/04/17**	**13**	**Tue**	Classification trees	Random forests	Bagging and boosting	
More modeling - catch up		**04/11/17**	**14**	**Tue**				
Delivering Results, “Big Data” - Hadoop and MapReduce	PDSwR 10, 11	**04/18/17**	**15**	**Tue**	Overview of Hadoop and MapReduce			
	Note: Additional web based readings and resources will be listed in Moodle							