Missouri University of Science & Technology
Information Science and Technology 5420

DATA METHODOLOGIES USING PYTHON
Spring Semester 2016 – Syllabus

Class Information

Class: Monday, Wednesday, and Friday 11:00-11:50
Class Room: Butler-Carlton Hall (BCH) 215

Contact Information

Instructor: Prof. Michael G. Hilgers Ph.D. CQF
Email: hilgers@mst.edu
Phone: 341-6484(office)
Office: Fulton Hall 106D
Office Hours: Monday, Wednesday & Friday (9:30-11:00) (Or by appointment: hilgers@mst.edu)

Course Information:

"By 2018, the United States alone will face a shortage of 140,000 to 190,000 people with deep analytical skills, as well as a shortfall of 1.5million data-savvy managers with the know how to analyze big data to make effective decisions."

Catalog Description

Python methodologies for manipulating, processing, cleaning, grouping, slicing, reshaping and summarizing information in data-intensive applications; managing files, scraping web pages, mining social media; describing, modeling, analyzing, and visualizing data. Tools include Pandas, NumPy, SciPy, and Matplo□b libraries.

Extended Description

Most analysis methods begins with the assumption that the data is "clean" meaning neatly organized and otherwise error-free. This course is concerned with the "dirty side" of data science. The raw data feeding the "big data" movement is messy. It is unstructured, incomplete, inconsistent, and erroneous. Cleaning data is not a canned process but requires specialized approaches based on careful inspections of the available raw data. Getting data ready for analysis needs methods in:

- Structuring data from a set of heterogeneous sources
- Knowing classification and conversion of data types and files
- Deciding among options in losing data due to erroneous and missing information
- Capturing data through remote sites via API's
- Scraping data from webpages
- Measuring the quality of resulting structures
- Distributing cleaned data

Python will be our tool of choice in applying these methods. It is a dirty job, but someone has to do it.

Course Prerequisites:

Programming – Prior experience with programming in an object-oriented language is required. Knowledge of file I/O, functions, looping, conditional flow, and data types will be used daily. No prior knowledge of Python is required.

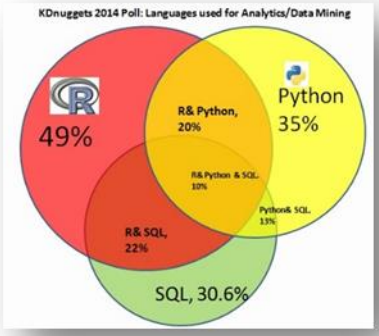
Calculus – Basic knowledge of a derivative and integral.

Statistics – Normal probability distributions, expected value, conditional expected value, mean, standard deviation, statistical inference

Textbooks:

We will cover most of the material in this book. The author solves many problems using MySQL. We will prefer Python in such situations. You will also need a good "Getting Started in Python" websites such as those found at python.org.

Software:



In this course, we will use Python as our language of choice. As can be seen from the diagram, R and Python dominate the data science world. R's numbers are bigger because it is associated with the heavy statistical analysis being used in the big data world. Python is a good workhorse that is simple in form, easy to learn, yet powerful in application.

Free sources of Python

Python can be downloaded at https://www.python.org/downloads/.

I used PyCharm as my IDE. It is available in some labs. It has a free download at: https://www.jetbrains.com/pycharm/download/#section=windows

Course Elements

Instructional Methods:

To achieve an understanding of the material, several techniques and methods will be used:

- I will lecture over foundational material. Typically, some form of notes will be provided, though I am not a strong supporter of PowerPoint in the classroom.
- Various sources of reading matter will be given including textbook, web, and personally developed notes. Please do the reading
- I like to work examples. Expect to spend some time studying data with software tools such as R.
- Analysis is not a spectator sport. Expect a homework problem or two most every night.

Assignments:

Homework/programming assignments will be made frequently. Please observe the following:

- Work is to be done individually unless otherwise specified. If you submit the work of another person as your own, you will receive a zero for the assignment and your name given to the Vice Provost of Undergraduate/Graduate studies.
- Homework is to be completed on the specified date and time.
- Late homework will NOT be accepted and will result in a zero for that assignment unless prior arrangement is made.

Course Content

The following should give you an idea of how we will proceed in the course. It is not ridged in that I will slow down rather than lose everyone or speedup if I am boring everyone.

Series and Data Frame Structures
Using Quandl's Python API we will look at time series objects and data frames. We will learn numerous ways to manipulate data frames.

Cleaning up a Mess

Using data concerning the Titanic passengers, we will examine ways to recover the "best" data from the dirty CSV file. This will create the opportunity to learn Python's extensive string manipulation capabilities.

Conversion between Data Types and File Types

We will learn how to convert data among file types such as CSV, JSON, and SQL. Facebook friendship networks will be used as a significant example.

Collecting Web Data

We will study regular expressions and how to implement them in Python to find HTML delimited data. We will also look at other web scraping tools.

RDBMS Cleaning Techniques

Using a mixture of MySQL and Python, we will see how to find anomalies in data bases and remove erroneous data.

Course Activity:

Given what is described about, the breakdown of activity is as follows:

Activity	Quantity	Course Percentage
Homework, Python Programs, Projects	5 (Approximately)	500 (100 pts each)
Midterm Examination or Project	1	250
Final Examination	1	250

Python programs: Practical applications of the concepts developed in class is extremely important. You will have a chance in this class to become fairly fluent in Python by the end of the semester.

Test: We will have two tests. The test will be drawn from lectures, examples, quizzes, and programs.

NOTE: It is very hard to build a course at this level. I may not give all assignments. I might give more. We will discuss and adjust the syllabus accordingly if this happens.

Grading Breakdown:

Grades will be based on total points, as defined below. There may be bonus points from time to time, which would be added to whatever category the bonus applies to. Boundaries for grades may be adjusted downward slightly, if deemed needed.

Grades:

A: 100% - 90%
B: 89% - 80%
C: 79% - 70%
D: 69% - 60%
F: Below 59%

Learning Objectives

	Communication Skills	Critical Thinking	Information Technology	Teamwork and Leadership
Develop basic skills needed to understand and manipulate the mathematical models forming the foundations of data analytics		X		
Learn how to use R to visualize large multidimenial data sets and explain	X	X	X	
Identify proper matematical model for a				
Perform Linear regression on a multi-dimension data set		X	X	
Be able to use nonparametric techniques		X	X	
Be able to use logistic models to analyze data sets		X	X	
Be able to use classification techniques to see patterns in data		X	X	
Be able to use clustering techniques to see patterns in data		X	X	
Be able to explain the role of business analytics in corporate environments	X			

COURSE POLICIES

Attendance:

Attendance is required, particularly as the assignments will be based on the important definitions and concepts presented in the lectures. You will likely want to ask questions. The class moves quickly and it is easy to fall behind and not get caught up. The more you miss class, the more material that will be foreign to you. If a student has missed an extended or excessive amount of classes or has failed to turn in multiple assignments, the instructor will send that student an Academic Alert. The alert will be emailed to the student and student's advisor. The student must meet with the instructor within three days or the instructor will send out another alert. If the student has not met with the instructor after the second alert, the instructor reserves the right to drop the student. If emergency circumstances arise, please contact the instructor soon to avoid penalties, and to try to catch up to the rest of the class.

Academic Integrity Statement

(http://registrar.mst.edu/academicregs/):

Violations of the University's academic code include, but are not limited to, possession of or use of unauthorized materials during quizzes or tests; providing unauthorized information to another student; or copying the work of another person. Violations may result in academic penalties in addition to receiving an "F" on the assignment in question. (See page 30 of S&T's "Student Academic Regulations" handbook for further details about student standards of conduct relative to the system's Collected Rules and Regulations section 200.010.) The most common attempt at dishonesty is submitting the program of another person with only some changes to deceive me. These are easy to recognize and not be tolerated.

Academic Alert System

(http://academicalert.mst.edu/):

S&T is committed to the success of its students by providing an environment conductive to teaching and learning. To ensure that every student takes full advantage of the educational opportunities and support programs on campus, the University has implemented an Academic Alert System, a web-based application. The purpose of the System is to improve the overall academic success of students by:

- Improving communication between students, instructors, and advisors;
- Reducing the time required for students to be informed of their academic status;
- Informing students of actions they need to perform in order to meet the academic requirements in the courses they are taking.

To assist you, I will initiate an academic alert for students who are not meeting academic course requirements through poor performance on assignments or poor attendance. When an alert is initiated, an email is immediately sent to the instructor, student, and advisor. You are encouraged to respond quickly to all academic alerts. If you fail to open the alert within one week, email notification is sent to your advisor.

Disability Support Services

(http://counsel.mst.edu/):

If you have a documented disability and anticipate needing accommodations in this course, you are strongly encouraged to meet with me early in the semester. You will need to request that the Disability Services staff send a letter to me verifying your disability and specifying the accommodation you will need before I can arrange your accommodation. If you have a disability that might require academic accommodations, please visit Disability Support Services in 204 Norwood Hall (341-4211; dss@mst.edu) very early in the semester.

Classroom Egress Maps

(http://registrar.mst.edu/links/egress/):

Please familiarize yourself with the classroom egress maps posted on line so you will know where emergency exits are located.

Cell Phones

Cell phones must be turned off will in the distance classroom.