The Ohio State University
The Max M. Fisher College of Business
Department of Management Sciences
BUSMGMT-7250: Data Management for Analytics Professionals
3 Credit hours, Fall Semester 2020

Instructor and contacts:

Dr. Dawit Mulugeta | (901) 517-0887 | mulugeta.14@osu.edu

Class Time and Location:

Lectures will be out on **Carmen** and **YouTube** on Wednesday &Thursday (4:30 PM-5:50 PM). The online classes will be complemented with three Saturday on-campus meetings (305 Gerlach Hall | 8:30 AM to 11:30 AM).

Instructor's Office Hours:

Thursday 5:30 PM to 7:00 PM and by appointment

Course Overview

Data and Data Management:

Data are being generated at a rapid rate, and we as a society are increasingly becoming a data-driven culture. Emerging technologies such as artificial intelligence, machine learning, internet of things, cloud computing, augmented analytics, big data analytics, Natural Language Processing and others are making headlines every day. Firms spend a big chunk of their budget on diverse tasks related to data. The demand for data scientist is skyrocketing. These developments signify that data are something to be understood, to be embraced, to be leveraged and to be revered. Although data is called in different names across different nations [Faamaumauga (Samoan), Paruru itepiri (Tahitian), Adatok (Hungarian), Datos (Spanish), Dados (Portugese) to name the few], it can be gold for everyone when it is processed.

Data are not monotonous entities. Data come in different variety (contents, shape and form), volume (size), velocity (speed) and veracity (authenticity). Along with comes the challenges of unravelling the hidden treasure within it as well as the prospects and the rewards of enhancing decision making processes. Predictive analytics, machine learning and other data science activities give enormous emphasis to data. The acquisition, the storage, the cleaning, the massaging, the aggregation, the characterization and the processing pieces are critical and monumental tasks. Such tasks often account for over 80% of the work in any data science projects. Data scientists both seasoned and beginners must acquire the necessary skills to understand the what, the where, the why, and the how aspects of data management and data processing.

Firms want to increase their ROI and allot their limited resources on product purchases, inventories, assortments, sales and marketing wisely. For this to happen they seek to understand customer's purchasing behavior including the amount, the frequency, the recency, and the trajectory of spends. Firms desire to know customer's product choices, willingness to pay, share of wallet, perceptions on products and services, pain-points and reasons for attrition. Firms also work hard to enhance loyalty and retention, to improve customer services in order to use the right marketing channel, to reduce risk, to increase profitability and to create win-win situations. All these cannot be done through gut-fillings, it requires coordinated efforts and large investments to acquire, to manage and to use diverse and large volume of data. Firms also require data-savvy analytic talent including predictive model builders, machine learning experts, busines intelligence specialists, business leaders, consultants, and others.

BUSMGMT-7250 aim to teach the art and the sciences of working with data. These include the understanding, the management, the manipulation, the massaging, the storage, the transformation, the aggregations and the extraction of hidden insights through analytics.

Note also that advances in information technologies and the increased digitization of business have led to an explosive growth in the amount of structured and unstructured data collected and stored in databases and other electronic repositories. Much, but certainly not all, of these data comes from operational business software (e.g., finance/accounting applications, Enterprise Resource Management (ERP), Customer Relationship Management (CRM), workflow and document management systems, surveillance and monitoring systems, and Web logs) and is often archived into vast data warehouses to become part of corporate memory. The result of this massive accumulation of data is that organizations have become data-rich yet still knowledge-poor. What can be learned from these mountains of data to improve decisions? How can an organization leverage its massive data warehouses for strategic advantage? A large number of methods with roots in statistics, informational retrieval and machine learning have been developed to address the issue of knowledge extraction from data sets - both small and large. The term "data-mining" refers to this collection of methods. These methods have broad applications; they have been successfully applied in areas as diverse as market-basket analysis of scanner data, customer relationship management, churn analysis, direct marketing, fraud detection, click-stream analysis, personalization and recommendation systems, risk management and credit scoring.

Course Objectives

1. Acquire a theoretical and a practical knowledge of contemporary data mining concepts and core analytics techniques.
2. Develop skills to manage, to analyze, to summarize, to report, and to present diverse data with the intent of telling compelling stories of actionable insights.
3. Gain hands-on experience in applying analytic techniques to practical real-word business problems using commercial data mining software.

Expected Outcomes

Upon completion of this course, students should be able to:

1. Become knowledgeable of contemporary data mining processes and analytic techniques to drive initiatives and strategies on various analytic projects.
2. Use existing data retrieval and manipulation tools and techniques to identify opportunities, to solve significant business problems and to extract actionable insights from data.
3. Fully appreciate the concept of data as a strategic resource, and understand how, when, where and why data mining can be used as a problem-solving technique.
4. Interpret, evaluate and describe the results of analytics and data mining works on a specific business issue.

Course Prerequisites

1. Good graduate standing and completion of an introductory course in probability and statistics.
2. Recognize the importance of analytics and data mining as a powerful information generation and decision-making tools.
3. Although students will use Python for data manipulation and analytics throughout the course, neither prior programming knowledge nor information technology background is required.

Text / Readings

Reference One (extensively used)
Title: Data Mining for Business Intelligence: Concepts, Techniques, and Applications, Second Edition.
Authors: Galit Shmueli, Nitin R. Patel, and Peter C. Bruce.
Publisher: John Wiley & Sons (2010).
ISBN: 9780470526828.
The book is available in digital form via the OSU library at:
https://learning.oreilly.com/library/view/data-mining-for/9780470526828/?ar

Reference Two (extensively used)
Title: Python: Data Analytics and Visualization.
Authors: Phuong Vo.T.H, Martin Czygan, Ashish Kumar, and Kirthi Raman.
Publisher: Packet Publishing (2017).
ISBN: 9781788290098.
The book is available in digital form via the OSU library at:
https://learning.oreilly.com/library/view/python-data-analytics/9781788290098/?ar

Reference Three (used in module 5)
Title: SQL: A Beginner's Guide, Third Edition, 3rd Edition.
Authors: Andy Oppel, Robert Sheldon.
Publisher: : McGraw-Hill (2008).
ISBN: 9780071548656.
The book is available in digital form via the OSU library at:
https://learning.oreilly.com/library/view/python-data-analytics/9781788290098/?ar

A set of articles, assignments, tutorials, data sets, lecture notes, and various supplementary materials will be made available through the course website on CARMEN as well as in You Tube. Please check the mentioned sites regularly to access newly posted materials, see when assignments are due and view reminders about the course.

Readings will be from the required text together with other supplementary materials. Some material will be covered only in the readings; others will be covered only in lecture which may depart from the text in either content or order. To maximize learning, classroom discussion and the amount of time spent on different topics will be adjusted according to the background and interests of the students.

Class Format

The teaching strategy of this course will be based primarily on lectures, in-class demonstrations, assignments, and classroom discussions.

Students can participate this course through CARMEN. Students are highly encouraged to visit the course site on CARMEN (https://carmen.osu.edu) regularly and print lecture materials in advance.

Throughout the course students are expected to bring their laptop into class. Each lecture will be complemented with associated and relevant work using the programming language Python. The laptop will be used to program with Python to manipulate data, to use the graphic interface, to develop various predictive and machine learning models, and to complete assignments.

As the field of data science and analytics encompasses several disciplines and are rapidly changing students are expected to read the selected reference materials and recommended readings for each topic together with the required textbooks.

Class Participation (10% of the final grade)

A portion of the final grade will be based on your class attendance and active participation, elements that are crucial to the success of class meetings. Attendance refers to punctual attendance. Your fellow students and I will expect you to come fully prepared to answer questions and discuss the assigned readings. Each individual is expected to actively and constructively contribute to class discussions. Good contributions transcend assigned readings and are inspired, timely, analytical, and relevant to the topics discussed. Students can also earn participation credit by drawing attention to related development, information and resources dealing with related topics. Your class participation grade will reflect my judgment of the quality and quantity of your contributions during the entire term.

Homework Assignments (30% of the final grade)

In addition to the reading requirements from the text and the supplementary materials, there will be six weekly homework assignments, spaced out over the course of the 7-weeks term. Each homework assignment is worth 5% of your final grade. They are designed to reinforce your understanding of the materials covered. Assignments are to be handed in on or before the class period of the due date. No late work is accepted. A limited amount of cooperation among students on homework and lab assignments is permitted. You may discuss with classmate's general solution strategies. However, everyone should independently do and turn in his/her own work.

Group Project (20% of the final grade)

Students will be randomly assigned into groups. Each group will consist of 3 to 6 students.
Tasks required from each group are as follows:
1) Identify and select data analytics topic from the list that will be provided or any topic that the team is interested to work on,
2) Write a one to two pages project proposal that includes:

- The background (what we know about the problem),
- The objective(s) (the purpose of the project),
- The materials and the methods (how the project will be executed and what analytic method will be used),
- The anticipated outcome (likely scenarios of outcomes) and
- How this will help achieve the objective(s) outlined.

3) Work on the project. This may include data extraction, data manipulation, data aggregation, data massaging, predictive modeling, and providing summary, and
4) Prepare a Power Point summary report. The report should include:

- Executive summary,
- Overview,
- Objectives,
- Materials and Methods,
- Results and discussion (include figures, tables, charts, demonstrations, etc.),
- Takeaways and suggestions for future line of work.

5) Present the project to the class on Saturday December 5 2020. Each team will have 15 to 20 minutes.

Exam (40% of the final grade)

There will be one online final exam. The exam is open-book and open-notes, and it will be held in accordance with Fisher Graduate Programs schedule during the final examination period on Wednesday December 9 (4:00 PM-5:45 PM).

The exam is designed to assess each student's (a) command of factual knowledge and concepts from the course; and (b) his or her ability to integrate and generalize these concepts and principles and apply them to new situations. The final exam must be taken during its scheduled time; make up exams will only be given for special and compelling cases, in accordance with University guidelines.

Software

The methods discussed in this class are computationally intensive and non-trivial; they cannot be performed using Excel. Fortunately, these methods have matured enough to the point where they are now implemented in commercial software.

Python will be the main programming language of the course. It is a free software widely used for data manipulation, statistical computing, machine learning, graphics, web applications and many more. It is supported by Phyton Software Foundation as well as large number of users. Python is versatile, easy to learn, has extensive libraries and is known to have large user base. It is one of the most extensively used software in the analytic community worldwide. Each lecture session will be accompanied by a demonstration of Python that focuses on specific tasks related to the discussion topic.

Grading

Class participation (10%).

Homework assignments (30%) .

Group Projects (20%).

Final exam (40%).

Feedback and Continuous Improvement

Students are strongly encouraged to visit with me in my office and/or use e-mail to ask questions, to share suggestions about any aspect of the course, or to clear up possible points of confusion. I will use your feedback to continuously improve and fine-tune the coverage levels and the teaching/learning processes. Please note that I may not always be able to make all of the changes suggested, but I will do my best to accommodate your suggestions.

Standards of Integrity and Conduct

Academic integrity is essential to maintaining an environment that fosters excellence in teaching, research, and other educational and scholarly activities. Each student in this course is expected to be familiar with and abide by the principles and standards set forth in The Ohio State University's Code of Student Conduct.

It is also expected that each student will behave in a manner that is consistent with the Fisher Honor Statement, which reads as follows:

As a member of the Fisher College of Business Community, I am personally committed to the highest standards of behavior. Honesty and integrity are the foundations from which I will measure my actions. I will hold myself accountable to adhere to these standards. As a future leader in the community and business environment, I pledge to live by these principles and celebrate those who share these ideals.

Safety and health requirements

All teaching staff and students are required to comply with and stay up to date on all University safety and health guidance, which includes wearing a face mask in any indoor space and maintaining a safe physical distance at all times. Non-compliance will be warned first and disciplinary actions will be taken for repeated offenses.

Disability Services

The university strives to make all learning experiences as accessible as possible. In light of the current pandemic, students seeking to request COVID-related accommodations may do so through the university's request process, managed by Student Life Disability Services. If you anticipate or experience academic barriers based on your disability (including mental health, chronic, or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.

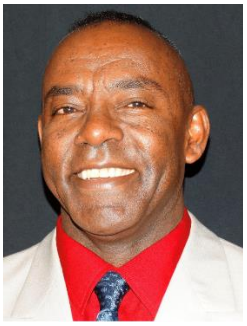
Module	Date	Lecture Topics & Readings
1	W (10/14)	**Part One: Course Introduction, Analytics and Current States of Affairs** • Tools, techniques and trajectories • Terms and definitions • Pillars of analytics • Data and analytics projects • **Readings:** *Reference One (Introduction & Chapter 1)* **Part Two: Introduction to Python** • **Readings**: *Reference Two (Module 1: Chapters 1 & 2)*
2	F (10/16)	**Part One: Data and Data Mining** • Definitions, types and sources • Dimension, consumption and trajectories • Point of sales, digital, panel, hierarchical, timeseries, invoice • Distribution • Data mining • Architecture, design, and governance • **Readings:** *Reference One (Chapter 2)* *1) Toward Scalable Systems for Big Data Analytics: A Technology Tutorial (https://ieeexplore.ieee.org/document/6842585)* *2) Big Data: The Management Revolution (https://hbr.org/2012/10/big-data-the-management-revolution)* **Part Two: Data Manipulation with Python** • **Readings**: *Reference Two (Module 1: Chapters 2 & 3; Module 2: Chapter 2)*
3	W (10/21)	**Part One: Data Extraction and Manipulation** • Rescaling, aggregation, zooming, filtering • Massaging (sorting, rearranging, transposing, merging) • Normalization. transformation • Imputation • Missing values and outlier detection and management • Visualization • **Readings**: *Reference One (Chapters 3 & 4)* **Part Two: Data Extraction and Manipulations with Python** • **Readings**: *Reference Two (Module 1: Chapters 3, 4 & 7)*

4	F (10/23)	**Part One: Data Exploration and Dimension Reduction** • Sampling and sampling methods • Variable selection, variable reduction • Data simulation • Correlation analysis • Principal Component Analysis (PCA) • Partitioning • **Readings**: *Reference One (Chapter 4)* **Part Two: Data Extraction and Manipulations with Python** • **Readings:** *Reference Two (Module 2: Chapter 3)*
	S (10/24)	**First Saturday: On-Campus Class Session** • 305 Gerlach Hall • 8:30 to 11:30 AM • Course review • Discussion of project progress, etc. • Q&A
5	W (10/28)	**Part One: Relational DBMS and SQL** • Principles and practices of DBMS • SQL: Data query, definition, control, manipulation and implementation • **Readings:** *Reference Three (Parts 1 & 2)* **Part Two: DBMS and SQL with Python** • **Readings:** *Reference Two (Module 2: Chapter 3)*
6	F (10/30)	**Part One: Emerging Data Technologies** • Big data analytics • Tools (Hadoop, Spark, etc.) • Web scrapping • **Readings:** *Reference (TBD)* **Part Two: Web scrapping with Python** • **Readings:** *Reference (TBD)*
7	W (11/4)	**Part One: Text Data and Text Analytics** • Files and sources • Feature extraction • Pre-and advanced text processing • **Readings:** *Reference (TBD)* **Part Two: Text analytics with Python** • **Readings:** *Reference (TBD)*
8	F (11/6)	**Part One: Predictive Model Development** • Standard processes and core principles • Model development and validation • Model deployment and documentation • **Readings:** *Reference One (Chapter 8)* **Part Two: Use of Python for predictive model development** • **Readings:** *Reference Two (Module 2: Chapters 1 & 4)*

9	W (11/11)	**Part One: Probability Theory and Practices** • Sample spaces, events and rules • Market Basket Analysis • Product Recommendation Systems • **Readings:** *Reference One (Chapter 13)* **Part Two: MBA & product recommendation system with Python** • **Readings**: *Reference Two (Module 2: Chapter 3)*
10	F (11/13)	**Part One: Linear Regression and its Applications** • Definitions, assumptions, diagnosis • Model parameters assessment • Training and validation • **Readings:** *Reference One (Chapter 6)* **Part Two: Linear Regression with Python** • **Readings:** *Reference Two (Module 1 Chapter 8; Module 2: Chapter 5)*
	S (11/14)	**Second Saturday: On-Campus Class Session** • 305 Gerlach Hall • 8:30 to 11:30 AM • Course review • Discussion of project progress, etc. • Q&A
11	W (11/18)	**Part One: Logistic Regression Theory and Practices** • Definitions, assumptions, diagnosis • Model parameters assessment • Training and validation • **Readings:** *Reference One (Chapter 10)* **Part Two: Logistic Regression with Python** • **Readings:** *Reference Two (Module 2: Chapter 6)*
12	F (11/20)	**Part One: Pricing Theory and Practices** • Framework for pricing and strategies • Pricing organizations • Pricing analytics • **Readings:** *Price- & Cross-Price Elasticity Estimation using SAS (https://support.sas.com/resources/papers/proceedings13/425-2013.pdf)* **Part Two: Pricing analytics with Python** • **Readings:** *Reference Two (Module 2: Chapter 5)*
13	W (11/25)	**Part One: Marketing Research (Forecasting, Conjoint Analysis, Lift Analysis)** • Univariate methods of forecasting • Conjoint product attributes, choice tradeoffs, relative utilities • Marketing campaigns and computation of lifts • **Readings:** *Reference One [Chapters 15,16,17 (forecasting); Chapter 5 (lift analysis), Conjoint (TBD)]* **Part Two: Marketing Research with Python** • **Readings:** *Reference (TBD)*

	F (11/27)	**No Class (Thanksgiving Break)**
14	W (12/02)	**Part One: Segmentation, Clustering, Decision Trees** • Purpose and advantage • Clustering types, rules and choices • Decision tree elements, rules, and steps • **Readings:** *Reference One (Chapters 7 & 14)* **Part Two: Clustering and decision trees with Python** • **Readings:** *Reference Two (Module 2: Chapters 7 & 8)*
15	F (12/04)	**Part One: Course Review Analytics Best Practices** • Course review • High level summary of topics covered **Part Two: Analytics Best Practices** • Practical tips in the area of analytics • The do's and the don'ts of analytics • **Readings:** *Reference Two (Module 2: Chapter 9)*
	S (12/05)	**Project Presentation** • 305 Gerlach Hall • 8:30 to 11:30 AM • Each project will have 15 to 20 minutes
	W (12/09)	**Final Exam**

About the Instructor



Dr. Dawit Mulugeta is a lecturer with the Fisher Business School of the Ohio State University. He is also a Vice President of Analytics and Risk Management with Wells Fargo. Earlier he had worked as a leader of advanced analytics with Cardinal Health, a healthcare services company that provides pharmaceutical and medical products and services in the United States and internationally. He had also analytics leadership stints with AutoZone, an automotive aftermarket retail company in Memphis, and Information Resources Incorporated (IRI), a leading market research and consulting firm in Chicago. Over many years he used contemporary analytics techniques and tools to work on diverse sets of business problems on customers, products, prices and services using big data. As a strategic thinker and a hands-on expert, he successfully implemented best in class decision support solutions. He is an applied statistician who loves the art and the sciences of big data, emerging technologies and analytics. He has published extensively and have made presentations at various analytics and scientific forums.

He had attended schools at Addis Ababa University in Ethiopia, Montana State University, and the University of Wisconsin in Madison. He lives in Powell Ohio with his wife, their four kids, and their Australian Shepherd Terrier mix puppy. His eldest two kids are current students at OSU. He loves to run, to play soccer, to coach soccer, to garden and to read. His association with OSU and with the Fisher Business school goes several years back when he served as a mentor to MBA or senior students. Go Bucks!