CSDA -1000 - Introduction to “BIG” Data Analytics

Room Location: ACW 203 (Accolade West)

Class Time: 9:30pm-5:30pm

Instructor: Matthew Tenney

Email: mtenney@yorku.ca

# COURSE DESCRIPTION:

The big data deluge has given birth to various new areas of data analysis and opened up new realms of opportunity for using data in decision making practices. This course offers an introduction to the field of data analysis practices with emphasis on four of the domains of the CAP® - Problem (Question) Framing, Analytics Problem Finding, Data and Methodology, and an understanding of the role of data analysis in making informed decisions. In this course you will also be provided with an overview of big data, big data analytics, and conducting a data analysis project responsibly as an end-to-end process with data exploration using Python.

You will actively participate through a combination of hands-on exercises, case studies and group work. You will be able to define the value proposition of data analysis by answering What? When? Where? Who? and How? questions when approaching a real-world problem. By the end of the course, you will be able to identify the skills necessary to develop core deliverables for stakeholders and design a data analysis project.

# LEARNING OUTCOMES:

Upon successful completion of this course, you will have acquired the knowledge, thought process and attitudes to:

1. Translate a problem statement into a business question that defines a purpose, scope, and requirements of a data analysis project.
2. Design and implement a data analysis project fitting real-world circumstances and stakeholder needs.
3. Differentiate data collection in terms of key characteristics and requirements.
4. Explore data sets to visually estimate the shape of key variables and to visually discover relationships between variables.
5. Use the Python programming language to do some data manipulation, data visualizations, and data analysis.

# COURSE FORMAT:

This course will be delivered in a full-day in-class format through eight (8) mandatory onsite sessions at York University. Onsite sessions shall consist of interactive plenary discussions, interspersed with small group discussions, followed by an afternoon practical lab session where concepts will be put into practice, and practical activities to facilitate understanding the concepts of data analysis, data analytics management, finding insights and presenting relevant insights to the business in a value added fashion each evening.

During the onsite sessions, students are expected to take notes and actively participate in discussions as each student brings unique experiences and perspectives that add value to the learning of all participants. Attendance is mandatory.

# COURSE OUTLINE:

|  |  |
| --- | --- |
| **Day 1 - Introduction** | **Topics** |
| Morning Lecture  (9am-10:30am) | The following major topics are introduced.   * Big Data * Data Analytics * Business Questions * Big Data and Analytics * Analytical Models * Statistical Methods |
| Lab  (11am-12pm; 1pm-2pm) | Introduction to Python for Data Analysis |
| Homework (2pm <) | Introduction to Programming Lab |
| Readings | Chen et al. 2014 – Introduction  Inform Chapter 1 & 2 |
|  |  |
| **Day 2 - Modeling the Analytic Problem** | **Topics** |
| Morning Lecture  (9am-10:30am) | The following major topics are introduced.   * Business Problem Definition * Analytics Problem Definition * Data Characteristics * Data Structure * Data Format * Data Granularity * Data Latency * Data Security * Variable Types * Data Visualization Basics |
| Lab  (11am-12pm; 1pm-2pm) | Introduction to Pandas |
| Homework (2pm <) | Pandas Worksheet |
| Readings | Chen et al. 2014 – Big Data Storage  Inform Chapter 3 |
|  |  |
| **Day 3 - Data Exploration and Discovery** | Topics |
| Morning Lecture  (9am-10:30am) | The following major topics are introduced.   * Data exploration * Data discovery * Measurement scales * Data shape * Univariate data * Bivariate data * Descriptive statistics * Correlation * Linear relationships |
| Lab  (11am-12pm; 1pm-2pm) | Introduction to Python for Data Analysis |
| Homework (2pm <) | Intermediate Data Analysis |
| Readings | Chen et al. 2014 – big data generation and acquisition |
|  |  |
| **Day 4 - Analytic Model Development** | Topics |
| Morning Lecture  (9am-10:30am) | The following major topics are introduced.   * Describe a range of different analytical modeling techniques * Statistical modeling * Data mining * Machine learning * Describe the Model Development Process CRISP-DM * Business Problem Framing * Data Collection and Evaluation * Data Preparation * Model Building * Model Evaluation * Model Deployment * Identify the structure and role of Programs, Projects and Services in terms of model development |
| Lab  (11am-12pm; 1pm-2pm) | Canadian Weather….is it colder? |
| Homework (2pm <) | Advanced Dataframes |
| Readings | Chen et al. 2014 – Big Data Storage  Inform Chapter 4 |
|  |  |
| **Day 5 - Research Methods** | Topics |
| Morning Lecture  (9am-10:30am) | The following major topics are discussed this class.   * Research concepts * Scientific method * Business questions * Critical thinking * Hypotheses * Sampling |
| Lab  (11am-12pm; 1pm-2pm) | Research Design |
| Readings | Chen et al. 2014 – Big Data Analysis |
|  |  |
| **Day 6 – Models Continued** | Topics |
| Morning Lecture  (9am-10:00am) | Feature Engineering and Model Evaluation |
| Lab  (10:30am-12pm; 1pm-2pm) | Geopandas Lab 1 & 2 |
| Homework (2pm <) | Liverpool Socio-demographic Case Study |
| Readings | Chen et al. 2014 – Big Data Applications |
|  |  |
| **Day 7 – Special Spatial Data** | Topics |
| Morning Lecture  (9am-10:00am) | Feature Engineering and Model Evaluation |
| Lab  (10:30am -12pm; 1pm-2pm) | Geopandas Lab 3 and 4 |
| Homework (2pm <) | Geopandas Worksheet |
| Readings | Chen et al. 2014 - Open Issues and Outlook |
|  |  |
| **Day 8 – Networks and Review** | Topics |
| Morning Lecture  (9am-10:00am) | The following major topics are discussed.   * Review, future discussions, and wrap up. |
| Lab  (10:30am-12pm; 1pm-2pm) | Walkability |
| Homework (2pm <) | Pandana Walkability Study in Toronto |
| Readings | None! |