

**Data Mining and Data Warehousing**

# General Course Information

|  |  |
| --- | --- |
| **Faculty/Title:** | Andres Fortino, Clinical Associate Professor |
| **Email:** | [agf249@nyu.edu](mailto:agf249@nyu.edu) |
| **Course title/number:** | Data Mining and Data Warehousing – MASY-GC3510- 203 |
| **Credits as applicable:** | 3 |
| **Semester/Year:** | Spring 2021 |
| **Session:** | 1 |
| **Number of Classes:** | 14 |
| **Dates/Day(s):** | **Tuesdays - 02/2/2021 – 05/4/2021** |
| **Time:** | 08:00am -- 10:35am EST/EDT |
| **No Class Dates:** | N/A |
| **Mode of Delivery:** | Online - Synchronous |
| **Special Notes:** |  |
| **Office Hours** | Email me to request an appointment – [agf249@nyu.edu](mailto:agf249@nyu.edu)  Appointments to be requested and scheduled after class starting at 10:45 AM EST/EDT |

# Course Description

In an increasingly competitive information age, data mining and data warehousing are essential in business decision-making. This course teaches students concepts, methods and skills for working with data warehouses and mining data from these warehouses to optimize competitive business strategy. In this course, students develop analytical thinking skills required to identify effective data warehousing strategies such as when to use outsource or in-source data services. Students also learn to Extract, Transform and Load data into data warehouses (the ETL process.) and use the CRISP approach to data mining to extract vital information for data warehouses. The course also teaches students how to secure data and covers the ethical issues associated with the uses of data and data models for business decisions.

# Course Prerequisites:

# Course Structure/Method

This course will be delivered Online - Synchronous , once a week on Tuesdays. Zoom is the remote instruction platform used at NYU. The class will encompass lectures, assignments, examples and demos, midterm and final exams, and a team project. All class content and assignments will be made available online via NYU Classes. Student should check the web site on a daily basis for any updates or announcements.

# Course Learning Outcomes

At the conclusion of this course, students will be able to:

* Translate business requirements into a well-constructed, normalized conceptual and logical data models
* Apply logical database design and the relational model
* Apply the CRISP model to conduct successful data mining
* Establish a successful ETL process to load a data warehouse
* Write basic SQL statements including some advanced SQL features
* Employ appropriate data governance principles to assure data quality and security

# Communication Policy

Credit students must use their NYU email to communicate. Non-degree students do not have NYU email addresses. NYU Classes course-mail supports student privacy and FERPA guidelines. All email inquiries will be responded to within 24 hours during Monday through Friday 5pm. Email sent on Saturday or Sunday will not be responded to until Monday. I will respond to you using NYU email.

Students have the opportunity to add their pronouns, as well as the pronunciation of their names, into Albert. Students can have this information displayed to faculty in Albert, NYU Classes, and other NYU systems. Students can also opt out of having their pronouns viewed by their instructors.

<https://www.nyu.edu/students/student-information-and-resources/registration-records-and-graduation/forms-policies-procedures/change-of-student-information/pronouns-and-name-pronunciation.html>

# Course Expectations

## Attendance

Students are expected to attend all classes. Excused absences are granted in cases of documented serious illness, family emergency, religious observance, or civic obligation. In the case of religious observance or civic obligation, this should be reported in advance. Unexcused absences from sessions may have a negative impact on a student’s final grade. Students are responsible for assignments given during any absence. Each unexcused absence may result in a student’s grade being lowered by a letter grade. A student who has three unexcused absences may earn a Fail grade.

University Calendar Policy on Religious Holidays <https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html>

Students who join the course during add/drop are responsible for ensuring that they identify what assignments and preparatory work they have missed and complete and submit those per the syllabus.

## Classroom Expectations

As graduate students, you are expected to conduct yourselves in a professional manner and engage and collaborate with your classmates in the Zoom meeting room. Here are our guiding principles:

* Dress as if you are in the Classroom.
* Keep your microphone muted unless asking a question or engaging in discussion.
* Check your video and audio when entering your class session.
* Think background, minimize distractions around you.
* Look into the camera instead of looking at the screen.
* Type quietly, mute if necessary.
* Don’t eat during a Zoom class session and refrain from engaging in any activity such as smoking, consuming alcohol, etc. that you would not engage in if the class was in person.

SPS classrooms are diverse and include students who range in age, culture, learning styles, and levels of professional experience. To maintain an inclusive environment that ensures all students can equally participate with and learn from each other, as well as receive feedback and instruction from faculty during group discussions in the classroom, all course-based discussions and group projects should occur in a language that is shared among all participants.

## Assignments

Students are expected to participate in each class session by offering their understanding of the subject, sharing ideas or discussing/commenting on other student’s comments. In addition, students must complete and submit all assigned homework on time. Late submission of homework will either not be accepted, or will result in a lower grade. Students are also expected to develop with and present a team project with other students, as well as take and pass a midterm exam and a final exam.

See full detail of expectations under “Assessment Strategy” below. Further information about specific assignments can also be found in the “Course Outline” section.

**Technology Policy**

Use of either PC or Mac desktop or laptop will be needed to access all course content/materials, notes, assignments and examples and demos. All class sessions require the use of Zoom.

## Class Participation

To receive full credit for class participation, you should attend all classes since much of the learning occurs during class lecture, presentation and class discussions. You must contribute and engage in class dialogue during every class session for the course. Please contact the instructor if you anticipate missing any part of the class. Please arrive on time so as not to disturb the flow of the lecture. Excessive lateness’s may result in lower overall grade.

Please contact the instructor if you anticipate missing any part of the class. Participation grades will be based on:

* Involvement in class discussions, dialogues and activities during each session
* Participation which demonstrates integration of reading, class work, relevance and application.
* Willingness to learn by accepting feedback, trying new skills and approaches, etc.
* Quality/quantity of providing effective and balanced feedback.

**Required and Recommended Material**

## Required

* The Data Warehouse Lifecycle Toolkit (2nd Edition) – Available through Amazon
  + **Authors** - Kimball, Ross, Thornthwaite, Mundy & Becker
  + **Publisher** – Wiley, 2008
  + **ISBN** - 978-0470149775
* Harvard Business School Publishers Cases
  + You must obtain the two classes used in class
  + Obtain a copy from <https://hbsp.harvard.edu/import/787977>
* Lantz, Brett. Machine learning with R: Expert techniques for predictive modeling. Packt Publishing Ltd, 2019. [L]
* Instructor will also provide session by session content, which will be posted online.

## Recommended Material

* Building the Data Warehouse (4th Edition)
* **Author** – W. H. Inmon
* **Publisher** – Wiley, 2005
* **ISBN** - 978-0764599446

# Software:

1. **Required JASP**, <https://jasp-stats.org/download/>
2. **Recommended– R, RStudio, RCommander,** <https://cran.r-project.org/>

**Assessment Strategy**

**Assignments** – Module Assignments – 2 Assignments (600 points, 60%)

Each of the major modules of the class will be concluded with an assignment in the form of an exercise to assure the student have mastered the material presented. Instructions for the assignments are posted to the class website. Late assignments will receive a 20% penalty in the grade. Assignments submitted more than one week late will receive no credit.

* Assignment 1 – Case Study 1 (Individual) - (200 pts, 20%)
  + Part A – Data Warehouse Justification (100 pts, 10%)
  + Part B – Data Warehouse Design (100 points, 10%)
* Assignment 3 – Case Study 2 (Team) - Call Center Performance (300pts, 30%)

**Labs** – Five graded Labs (250 points, 25%). In the data mining second half of the class there will be a lab every other week. The answers to the labs will be entered in the appropriate Quiz in the NYU Classes class website. They are due one week the class. Lab 6 (optional) score can substitute for the lowest score of any of the other labs.

* + Lab 1 - Designing a Data Warehouse Data Schema (50 pts, 5%)
  + Lab 2 - Framing Analytical Questions (50 pts, 5%)
  + Lab 3 - Data Preparation (50 pts, 5%)
  + Lab 4 - Descriptive Data Mining (50 pts, 5%)
  + Lab 5 - Predictive Data Mining (50 pts, 5%)
  + Lab 6 - Practice Final Exam (optional) (50 pts, 5%)

**Final** – There will be a short 1 hour in-class final. (50 points, 5%). A practice final exam will be made available for you to practice taking the final exam: Lab 6.

**Review Quizzes** (100 points, 10%)**:**

There are12 required REs (Reflection Exercises), 5 points each for taking them, not based on the score. This is not an exercise to measure what you know but to assist you to transfer knowledge from short term memory to long term memory. **In the past students who used these exercises got as much as a 30% increase in their final exam grade.** The quizzes are open online for a whole week and they are timed to maximize knowledge transfer. So be sure to take each quiz when it is available.

# NYUSPS Policies

NYUSPS policies regarding the Family Educational Rights and Privacy Act (FERPA), Academic Integrity and Plagiarism, Students with Disabilities Statement, and Standards of Classroom Behavior among others can be found on the NYU Classes Academic Policies tab for all course sites as well as on the University and NYUSPS websites.   
Every student is responsible for reading, understanding, and complying with all of these policies.

The full list of policies can be found at the web links below:

University: <http://www.nyu.edu/about/policies-guidelines-compliance.html>

NYUSPS: <https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1>

# Center for Student Accessibility

If you are a student who is requesting accommodations, please contact New York University’s Moses Center for Students Accessibility (CSA) at 212-998-4980 or mosescsa@nyu.edu. You must be registered with CSA to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csa. The Moses Center is located at 726 Broadway on the 3rd floor.

# Health and Wellness

To access the University's extensive health and mental health resources, contact the NYU Wellness Exchange. You can call its private hotline (212-443-9999), available 24 hours a day, seven days a week, to reach out to a professional who can help to address day-to-day challenges as well as other health-related concerns.

# Academic Integrity and Plagiarism Policy

All students are expected to be honest and ethical in all academic work. This trust is shared among all members of the University community and is a core principle of American higher education. Any breaches of this trust will be taken seriously. A hallmark of the educated student and good scholarship is the ability to acknowledge information derived from others. Students are expected to be scrupulous in crediting those sources that have contributed to the development of their ideas.

Plagiarism involves borrowing or using information from other sources without proper and full credit. Students are expected to demonstrate how what they have learned incorporates an understanding of the research and expertise of scholars and other appropriate experts; and thus recognizing others' published work or teachings—whether that of authors, lecturers, or one's peers—is a required practice in all academic projects.

Students are subject to disciplinary actions for the following offenses which include but are not limited to: Cheating; Plagiarism; Forgery or unauthorized use of documents; False form of identification

Use the link below to read more about Academic Integrity Policies at the NYU School of Professional Studies. [Academic Policies for NYU SPS Students](https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1)

# NYU School of Professional Studies Graduate Grading Scale

Grading for graduate programs is by letter grade: A, A-, B+, B, B-, C+, C, C-, and F. For NYUSPS’s complete graduate grading policies, including criteria for a grade of incomplete, taking a course on a pass/fail basis, and withdrawing from a course, see: <https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Letter** | **%** | **GPA** | **Descriptions** | **Definitions** |
| A | 95-100 | 4.0 | Exceptional | Demonstrates exceptional mastery of all learning outcomes of the course and thorough and complete understanding of all concepts. |
| A- | 90-94 | 3.7 | Excellent | Demonstrates highly competent mastery of all learning outcomes of the course and strong understanding of all concepts. |
| B+ | 87-89 | 3.3 | Very Good; exceeds course standards | Demonstrates mastery of all learning outcomes of the course and understanding of core concepts. |
| B | 83-86 | 3.0 | Good; meets course standards | Demonstrates mastery of some learning outcomes; understanding of some core concepts could be improved. |
| B- | 80-82 | 2.7 | Somewhat Satisfactory; meets some course standards and requires improvement | Demonstrates basic understanding of some learning outcomes; improved understanding of all core concepts is needed. |
| C+ | 77-79 | 2.3 | Less than Satisfactory; requires significant improvement | Demonstrates partial understanding of all learning outcomes and core concepts; requires significant improvement. |
| C | 73-76 | 2.0 | Unsatisfactory; requires substantial improvement | Demonstrates partial understanding of some learning outcomes and core concepts; requires substantial improvement. |
| C- | 70-72 | 1.7 | Unsatisfactory; requires extensive improvement | Demonstrates poor understanding of all learning outcomes and core concepts; requires extensive improvement. |
| F | Below 70 | 0.0 | Fail | Demonstrates minimal to no understanding of all key learning outcomes and core concepts; work is unworthy of course credit towards the degree. |

# Course Outline

**Start/End Dates: 02/2/2021 – 05/4/2021**

**Readings should be completed within the week.**

## Week 1, Session 1, **02/02/21** - Introduction to Data Warehousing

Module 1: What are Data Warehousing and Data Mining?

We answer the question: *What is a data warehouse?*

* Introduction to Data Warehousing
* Relationship of Data Mining and Data Warehousing
* What is a Data Warehouse?
* Data Warehousing ROI
* DSS - Decision Support Systems
* Operational vs. Analytical Systems
* Evolution of DSS and Data Warehousing
* OLTP - Online Transaction Processing
* Characteristics of a Data Warehouse
* What is a Data Mart? Creating a Data Mart
* Data Comparison Chart
* OLAP - Online Analytical Processing
* **Assignments (due one week from today):**
  + **Reading:** Chapter 1 (from both Data Warehouse Lifecycle Toolkit, and Building the Data Warehouse).

## Week 2, Session 2, **02/09/21** - Planning and Building the Data Warehouse

Module 2: Data Warehousing 1 - Planning the Data Warehouse

We answer the question: *Why do we build and use a data warehouse?*

* Planning & Building the Data Warehouse
* Sponsorship and Cost Justification
* Project Prerequisites
* Barriers, Challenges and Risks
* Preparing for Implementation
* Developing the Data Warehouse
* SDLC Methodologies - Waterfall vs. RUP Approach
* Planning & Project Management
* Analysis
* Logical & Physical Design
* Implementation and Deployment
* Operations
* **Assignments due:**
  + **Reading:** Chapter 1, 2 (The Data Warehouse Lifecycle Toolkit)

## Week 3, Session 3, **02/16/21** - Data Warehouse Design

Module 3: Data Warehousing 2- Data Warehouse Design

We answer the question: *How are data warehouses designed?*

* Data Warehouse Design
* Drivers for Multi-Dimensional Analysis
* Limitations of Relational Models
* The Data Cube
* What is dimensional modeling?
* Advantages of Dimensional Models
* Logical and Physical Design
* Data Normalization
* Benefits and Drawbacks of Data Normalization
* De-Normalizing of Data
* Characteristics of a Data Warehouse
* Subject Oriented, Integrated, Time Variant, Non-Volatile
* The Star Schema
* **Assignments due:**
  + **Assignment 1A - Data Warehouse Case Due**
  + **Reading:** Chapter 6 (The Data Warehouse Lifecycle Toolkit)

## Week 4, Session 4, **02/23/21** - Components of a Data Warehouse

Module 4: Data Warehousing 3 - Components of a Data Warehouse

We answer the question: *How are data warehouses built?*

* Data Warehouse Schemas
* Dimensions and Dimension Tables
* Facts and Fact Tables
* The Star Schema
* The Snowflake Schema
* Degenerate and Junk Dimensions
* The Data Warehouse Bus Architecture
* Conformed Dimensions and Standard Facts
* Data Granularity
* Changing Dimensions
* Components of a Data Warehouse
* Source Systems, Staging Area, Presentation, Access Tools
* Building the Data Matrix
* The Four Steps Process
* Multiple Fact Tables in a single Data Mart
* Chain, Heterogeneous, Transaction/Snapshot & Aggregate Facts
* Fact and Dimension Table Detail
* Identifying Source for each Fact & Dimension
* Mapping from Source to Target
* **Assignments due:**
* **Reading:** Chapter 4, 6 and 7 (The Data Warehouse Lifecycle Toolkit)

## Week 5, Session 5, **03/02/21** - The ETL Process

Module 5: Data Warehousing 4 - Loading the Data Warehouse - ETL

We answer the question: *How are data warehouses loaded with data?*

* The ETL Process
* Extracting the Data into the Staging Area
* The Challenge of Extracting from Disparate Platforms
* Full vs. Incremental Extracts
* Detecting Changes to Data
* Transforming the Data
* Complexity of Data Integration
* Dealing with Missing & Dirty Data
* Data Transformation Tasks
* Loading the Data
* Timing and Job Control of Data Loads
* **Assignments due:**
  + **Reading:** Chapter 9 (The Data Warehouse Lifecycle Toolkit)
  + **Assignments:** Lab 1 - Create a logical database model using data normalization rules

**Week 6, Session 6,** 03/09/21 **-**

Module 6: Data Warehousing 5 - Extracting for Data Mining- SQL

We answer the question: *How do we extract data from warehouses for data mining?*

* **Assignments due:**
  + **Reading:** Chapter 9 (The Data Warehouse Lifecycle Toolkit)
  + **Assignments:** Assignment 1B - Data Warehouse Case Due

## Week 7, Session 7, **03/16/21** – Introduction to Data Mining

Module 7: Data Mining 1 - Data Mining, CRISP-DM and Framing Questions

We answer the question: *What is Data Mining?*

* Data Mining and Concepts
* What is Data Mining Good For?
* Statistics, Artificial Intelligence & Machine Learning
* Data Mining Examples and Tools
* Connection between Data Mining and Data Warehousing
* Retrospective Reporting vs. Predictive
* Data Mining Applications
* Data Mining vs. Statistics vs. OLAP
* Keys to commercial success of Data Mining
* **Assignments due:**
  + **Reading:** Chapter 9 (The Data Warehouse Lifecycle Toolkit)
  + **Assignment:** Lab 2  Framing Questions

## Week 8, Session 8, **03/23/21** - Data Preparation

Module 8: Data Mining 2 - Data Preparation

We answer the question: *How do we prepare data for data mining?*

* Data Preparation
* **Assignments due:**
* **Reading:** Chapter 8 p353-357 (The Data Warehouse Lifecycle Toolkit)

## Week 9, Session 9, **03/30/21** – Descriptive Data Mining

Module 9: Data Mining 3 – Descriptive Data Mining

We answer the question: *What is Descriptive Data Mining?*

* Data Mining Statistical Techniques (Sampling, Regression & Decision Trees)
* Data Mining Techniques
* Terminology
* **Assignments due:**
  + **Reading:** Chapter 2, 3, 5 (The Data Warehouse Toolkit)
  + **Assignments:** Lab 3 Data Preparation

## Week 10, Session 10, **04/06/21** – Predictive Analytics

Module 10: Data Mining 4 – Predictive Analytics and Linear Regression

We answer the question: *What is Predictive Data Mining?*

* Linear Regression
* Linear Regression with multiple independent variables
* **Assignments due:**
  + **Reading:** Chapter 15 (The Data Warehouse Toolkit)

## Week 11, Session 11, **04/13/21** Analytic Models – Logistical Regression

Module 11: Data Mining 5 – Analytic Models – Logistical Regression

We answer the question: *What is Supervised Machine Learning?*

* Clustering, Segmentation and Nearest Neighbor Techniques
* Decision Trees
* **Assignments due:**
  + **Reading:** Online web research and reading
  + **Assignment:** Lab 4 - Descriptive Data Mining

## Week 12, Session 12, **04/20/21** - Data Mining 6: Decision Trees and Clustering

Module 12: Data Mining 6 – Analytic Models – Decision Trees and Clustering

We answer the question: *What is Unsupervised Machine Learning?*

* Data Mining Techniques – Part 2
* Supervised vs. Unsupervised Data Mining
* Classification and Segmentation
* Tree Inductions
* Entropy and Information Gain
* The ID3 and C4.5 classifier process
* Clustering
* Co-occurrence grouping
* Examples using Weka
* **Assignments due:**
  + **Reading:** Online web research and reading
  + **Assignment:** Assignment 2 Titanic Kaggle Competition Due

## Week 13, Session 13, **04/27/21** - Data Mining Techniques II

Module 13: Data Mining 7 – Model Building and Maintenance

We answer the question: *How do we build and maintain data models?*

* Data Models
* Model Maintenance
* **Assignments due:**
  + **Reading:** Online web research and reading
  + **Assignment:** Lab 5 - Predictive Data Mining

## Week 14, Session 14, **05/04/21** – Weka

Module 14: Exam

We answer the question: *Can you pass the INFORMS CAP Certification exam?*

**Final Exam**

* **Assignments due:**
  + **Assignment:** Lab 6 Final Exam Pratice (optional)
  + **Data Warehouse Team Project:** Assignment 3 Team Call Center Case due

**Couse Schedule**

Table

Description automatically generated

***At the discretion of the faculty, the syllabus may be modified to better meet the needs of the students and to achieve the learning outcomes established in the syllabus******.***