Knight Foundation School of Computing and Information Sciences

Course Title: Introduction to Data Science **Date:** 6/2/2024

Course Number: CAP 2757

Number of Credits: 3

Subject Area: Applications
Catalog Description: Fundamental data science lifecycle topics with key concepts in data ethics, governance, applied statistics, and computing with hands-on experience to apply knowledge in real-world scenarios.
Textbooks: Data Science from Scratch, 2nd Edition by Joel Grus. Released May 2019. Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041139.
References (for further reading): Python for Data Analysis, 3rd Edition by Wes McKinney. Released August 2022. Publisher(s): O'Reilly Media, Inc. ISBN: 9781098104030.
Prerequisites Courses: COP 2047 – Python Programming I or Advisor's Permission
Corequisite Courses: COP 3045 – Computational Thinking

Type: Core Course for BS in Data Science; Elective for CS and IT Majors.

Prerequisites Topics:

1. Basic Programming Concepts (variables, data types, loops, conditionals, and fundamental algorithms)
2. Functions and Module Programming (writing and invoking functions)
3. Fundamental Statistics (concepts like mean, median, mode, standard deviation, and basic probability)
4. Python Libraries and Tools (familiarity with Python's standard libraries)

Course Outcomes:

1. **Explain** the data science lifecycle and its role in various industries.
2. **Experiment** with data acquisition, management, and integration.
3. **Apply** statistical techniques for data analysis and hypothesis testing.
4. **Utilize** machine learning algorithms for model development and evaluation.
5. **Create** effective data visualizations and communicate insights derived from data.
6. **Explore** ethical considerations, governance, and privacy concerns in data science projects.
7. **Apply** fundamental computing concepts, including programming, data structures, and algorithms.
8. **Develop** a real-world data science projects by integrating and applying knowledge and skills acquired throughout the course.

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

Association between Student Outcomes and Course Outcomes

BS in Computing: Student Outcomes Graduates of the program will have an ability to:	Course Outcomes
1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.	1,7,8
2) Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.	4,7,8
3) Communicate effectively in a variety of professional contexts.	5,8
4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.	6
5) Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.	1,8

Program Specific Student Outcomes	
6) Apply theory, techniques, and tools throughout the data science lifecycle and employ the resulting knowledge to satisfy stakeholders' needs. [DS]	1,2,3,4,5,6,7,8

Assessment Plan for the Course and how Data in the Course are used to assess Student Outcomes

Student and Instructor Course Outcome Surveys are administered at the conclusion of each offering, and are evaluated as described in the School's Assessment Plan: https://abet.cis.fiu.edu/

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

Outline

Topic	**Number of Lecture Hours** (Total: 37.5 hours = 15 weeks * 2 lectures/week * 1.25 hrs/lecture)	Outcome
1. Introduction to Data Science 1.1. Definition and history of data science 1.2. Data science lifecycle overview 1.3. Roles and responsibilities in data science teams 1.4. Importance of data science in various industries 1.5. Data Acquisition and Representativeness	3.75	1
2. Types of data sources (structured, unstructured, semi-structured) 2.1. Data collection methods and tools 2.2. Data cleaning and preprocessing 2.3. Working with missing or incomplete data 2.4. Data quality and representativeness 2.5. Sampling techniques	3.75	2
3. Basic concepts/review in: 3.1. Linear algebra 3.2. Probability 3.3. Optimization	3.75	3
4. Data Management 4.1. Data storage systems (databases, data warehouses, data lakes) 4.2. Data formats and file types 4.3. Data indexing and querying 4.4. Data integration techniques	3.75	2,4
5. Data Preparation and Integration 5.1. Data cleaning and preprocessing 5.2. Handling missing data 5.3. Feature engineering 5.4. Data transformation and normalization	3.75	2,3,4

6. Data Analysis 6.1. Measures of central tendency 6.2. Measures of dispersion 6.3. Data distributions 6.4. Descriptive statistics 6.5. Exploratory data analysis 6.6. Inferential statistics 6.7. Hypothesis testing and confidence intervals	3.75	2,3,4
7. Model Development and Deployment 7.1. Overview of machine learning and its applications 7.2. Fundamentals of supervised learning (regression, classification) 7.3. Model evaluation and validation 7.4. Deployment and monitoring of models	3.75	2,3,4,7
8. Data Visualization and Communication 8.1. Principles of effective data visualization 8.2. Visualization tools and libraries 8.3. Storytelling with data 8.4. Reporting and presentation best practices	3.75	5,7
9. Introduction to Data Ethics, Governance, and Privacy 9.1. Ethical considerations in data science 9.2. Algorithmic fairness and bias 9.3. Data privacy and security 9.4. Data stewardship (protection and preservation)	3.75	6,8
10. Project-based Learning 10.1. Students work on a project that incorporates an application area and requires integration and application of knowledge and skills acquired in earlier course work.	3.75	7,8

Performance Measures for Evaluation

All assignments are assigned through the Canvas course site. Please note that the deadlines are strictly enforced. For example, if the deadline is 11:59 PM, any assignment submitted after this time is considered late. It is also each student's responsibility to submit correct files and ensure the submission is successful before the deadline (please double check your Canvas submissions). If you are unable to submit your assignment through Canvas, send a copy of your assignment to your instructor before the stated deadline. There will be three exams and each exam will be cumulative with an emphasis on the most recently covered material. Please note that every student is required to be physically present to take the exams with their own laptop. Exam details will be posted on the Canvas course site (https://canvas.fiu.edu).

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

Assignment	Points Each	Total Points	Percentage of Final Grade
Quizzes (11-Drop-1)	10	100	10%
Homework Assignments (2)	100	200	20%
Exam 1	200	200	20%
Exam 2	200	200	20%
Class Project	300	300	30%
		TOTAL	100%

Letter Grade Distribution Table

Letter	Range%	Letter	Range%	Letter	Range%
A	93 or above	B	82 - 85.9	C	70 - 73.9
A-	90 - 92.9	B-	78 - 81.9	D	60 - 69.9
B+	86 - 89.9	C+	74 - 77.9	F	less than 60

Description of Possible Homework Activities

Homework 1: Data Exploration and Preparation
Description: This assignment aims to help students apply the skills they've acquired from the beginning of the course through Topic 5, focusing on understanding, preparing, and integrating data.

Tasks:

1. **Dataset Selection and Acquisition:** Choose a dataset (preferably real-world) that contains both structured and unstructured data elements. Justify the dataset selection in the context of a hypothetical business problem.
2. **Data Cleaning and Preprocessing:**
    - Identify and handle missing or incomplete data.
    - Use appropriate techniques to clean and preprocess the data.
    - Validate the quality and representativeness of the cleaned data.
3. **Data Management:**
    - Store the dataset in an appropriate format. Discuss why you chose this format.
    - Describe the indexing method you'd use to quickly retrieve data.
4. **Linear Algebra and Probability Application:** Using the cleaned dataset, perform the following:
    - Identify correlations between variables using linear algebra concepts.
    - Compute the probability of a certain event or category in the dataset.
5. **Data Integration:** If you've chosen multiple datasets or if the dataset has multiple sources, integrate them using appropriate methods.

Deliverables: A report containing a description of the dataset, business problem, data

preparation methods, results from linear algebra and probability tasks, and any insights or challenges faced.

Description of Possible Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)	Weight
Data Selection	Dataset is highly relevant and perfectly aligns with the hypothetical business problem, with a well-justified selection	Dataset is relevant and mostly aligns with the hypothetical business problem, with a good justification for the selection	Dataset is somewhat relevant and somewhat aligns with the hypothetical business problem, with a basic justification for the selection	Dataset is slightly relevant but does not align well with the hypothetical business problem, with a weak justification for the selection	Dataset is not relevant and does not align with the hypothetical business problem, with no or incorrect justification for the selection	10%
Data Cleaning and Preprocessing	Comprehensive and efficient data cleaning demonstrating mastery, with validation of data quality and representativeness	Good data cleaning and preprocessing with a reasonable validation of data quality and representativeness	Basic data cleaning and preprocessing with minimal validation of data quality and representativeness	Insufficient data cleaning and preprocessing with inadequate validation of data quality and representativeness	No or incorrect data cleaning and preprocessing with no validation of data quality and representativeness	15%
Data Management	Perfectly chosen format and indexing method with clear and well-articulated justifications	Well-chosen format and indexing method with good justifications	Adequately chosen format and indexing method with basic justifications	Poorly chosen format and indexing method with weak justifications	Incorrectly chosen format and indexing method with no or incorrect justifications	15%
Linear Algebra and Probability Application	Excellent application of linear algebra and probability concepts to identify correlations and compute probabilities, showcasing deep understanding	Good application of linear algebra and probability concepts to identify correlations and compute probabilities	Basic application of linear algebra and probability concepts to identify some correlations and compute probabilities	Limited application of linear algebra and probability concepts with insufficient identification of correlations and computation of probabilities	No or incorrect application of linear algebra and probability concepts, failing to identify correlations and compute probabilities	15%
Data Integration	Flawlessly integrated data from multiple sources using best practices, demonstrating deep understanding	Well-integrated data from multiple sources using good practices	Moderately integrated data from multiple sources using basic practices	Poorly integrated data from multiple sources using inadequate practices	No or incorrect data integration, failing to properly combine data from multiple sources	25%
Report Quality	Report is exceptionally well-written, providing a detailed description and insightful analysis, with clear documentation of insights and challenges faced	Report is well-written, providing a good description and analysis, with documentation of most insights and challenges faced	Report is adequately written, providing a basic description and analysis, with some documentation of insights and challenges faced	Report is poorly written, providing a limited description and analysis, with little documentation of insights and challenges faced	Report is not written or is incorrectly written, providing no or incorrect description and analysis, with no documentation of insights and challenges faced	20%

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

Homework 2: Comprehensive Data Analysis and Visualization

Description: This assignment integrates the latter topics of the course, emphasizing data analysis, modeling, visualization, and initial steps into ethical considerations.

Tasks:

1. **Data Analysis:** Using the dataset from the mid-term assignment (or a new one):
    - Conduct an exploratory data analysis (EDA).
    - Use inferential statistics to test a hypothesis related to the data.
    - Identify and interpret the measures of central tendency and dispersion for key variables.
2. **Model Development:**
    - Based on the insights from the EDA, choose a suitable machine learning approach (regression or classification).
    - Train and evaluate the model. Discuss the model's performance using appropriate evaluation metrics.
3. **Data Visualization:**
    - Develop at least three different visualizations that showcase key findings or insights from the data.
    - Ensure that these visualizations adhere to the principles of effective data visualization discussed in class.
4. **Ethical Considerations:** Reflect on potential ethical issues related to your dataset, analysis, or findings. Consider aspects like biases in the data, potential misuse of the information, and privacy concerns.

Deliverables: A comprehensive report detailing your EDA, hypothesis testing, machine learning model, visualizations, and ethical reflections. This should also include a section discussing how you might deploy and monitor the model in a real-world scenario.

Description of Possible Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)	Weight
Data Analysis	Demonstrates in-depth understanding and mastery in EDA, measures, and statistics, showcasing deep insights and comprehensive analysis	Demonstrates good understanding in EDA, measures, and statistics, showcasing substantial insights and analysis	Demonstrates average understanding in EDA, measures, and statistics, showcasing some insights and analysis	Demonstrates below-average understanding in EDA, measures, and statistics, showcasing limited insights and analysis	Demonstrates poor understanding in EDA, measures, and statistics, showcasing no or incorrect insights and analysis	20%
Model Development	Model is highly effective, perfectly justified, and validated with excellent performance metrics	Model is effective, well-justified, and validated with good performance metrics	Model is moderately effective, somewhat justified, and validated with average performance metrics	Model is slightly effective, poorly justified, and validated with below-average performance metrics	Model is not effective, unjustified, and not validated with poor or incorrect performance metrics	20%

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

Data Visualization	Develops exceptional, clear, and insightful visualizations that perfectly adhere to best practices	Develops good, clear, and somewhat insightful visualizations that mostly adhere to best practices	Develops average visualizations that somewhat adhere to best practices and showcase some insights	Develops below-average visualizations that barely adhere to best practices and showcase limited insights	Develops poor visualizations that do not adhere to best practices and showcase no or incorrect insights	20%
Ethical Considerations	Demonstrates a deep understanding of potential ethical issues with actionable insights and comprehensive reflection	Demonstrates a good understanding of potential ethical issues with substantial insights and reflection	Demonstrates an average understanding of potential ethical issues with some insights and reflection	Demonstrates a below-average understanding of potential ethical issues with limited insights and reflection	Demonstrates a poor understanding of potential ethical issues with no or incorrect insights and reflection	20%
Report Quality and Real-world Application	Report is exceptionally well-written, providing a detailed description and insightful analysis, with a well-thought-out plan for real-world deployment and monitoring	Report is well-written, providing a good description and analysis, with a reasonable plan for real-world deployment and monitoring	Report is adequately written, providing a basic description and analysis, with a simple plan for real-world deployment and monitoring	Report is poorly written, providing a limited description and analysis, with a vague or incomplete plan for real-world deployment and monitoring	Report is not written or is incorrectly written, providing no or incorrect description and analysis, with no plan for real-world deployment and monitoring	20%

Class Project: Real-World Data Analysis and Insights

Description: For the final project, students will analyze a real-world dataset of their choice (or provided by the instructor) and use the skills and knowledge which they have acquired throughout the course to extract insights, develop a predictive or descriptive model, and effectively communicate their findings. It is recommended that the students follow a set of steps in order to produce a high-quality project:

1. **Data Selection:** Choose a dataset that is sufficiently complex and large enough to warrant meaningful analysis but manageable given the time constraints of the course.
2. **Data Cleaning and Preprocessing:** Demonstrate the ability to process and prepare the data for analysis, including handling missing values, outliers, and any transformations necessary.
3. **Exploratory Data Analysis (EDA):** Conduct a thorough EDA to understand the dataset's characteristics and patterns.
4. **Statistical Analysis:** Apply descriptive and inferential statistics to understand the data distribution and test hypotheses.
5. **Model Development:** Develop a machine learning model based on the problem statement (e.g., prediction, classification).
6. **Data Visualization:** Design effective visualizations to communicate your insights and findings.
7. **Ethical Consideration:** Evaluate and address any ethical concerns related to the data, analysis, or potential implications of the model's deployment. Consider aspects like data privacy, potential biases in the data, fairness, and the consequences of false positives or negatives.

Knight Foundation School of Computing and Information Sciences
CAP 2757 Introduction to Data Science

8. **Report and Presentation:** Compile your analysis, methodology, and insights in a comprehensive report. Additionally, prepare a presentation to communicate your findings to the class.

Description of Possible Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)	Weight
Data Selection	Dataset is highly relevant, of appropriate complexity, and allows for in-depth analysis.	Dataset is relevant and of appropriate complexity for analysis.	Dataset is somewhat relevant and may be too simple or too complex for the course level.	Dataset has limited relevance and is not ideal for meaningful analysis.	Dataset is not relevant or inappropriate for analysis.	10%
Data Cleaning and Preprocessing	Comprehensive and efficient data cleaning. Demonstrates exceptional skills in preprocessing.	Effective data cleaning and preprocessing with minor oversights.	Adequate data cleaning, but with some significant gaps or errors.	Limited data cleaning and preprocessing. Numerous errors.	Inadequate or no data cleaning.	15%
Exploratory Data Analysis	In-depth EDA with thoughtful insights. Shows mastery in understanding the data.	Good EDA with some meaningful insights.	Adequate EDA but misses out on significant patterns.	Limited EDA with few insights.	Minimal or no EDA.	10%
Statistical Analysis	Comprehensive statistical analysis with appropriate tests and valid conclusions.	Good statistical analysis with minor errors in interpretation.	Some useful statistical tests applied but with errors.	Few statistical methods applied, significant misunderstandings.	Minimal or no statistical analysis.	10%
Model Development	Model is highly effective, well-justified, and validated. Demonstrates mastery in modeling.	Model is effective with minor issues in justification or validation.	Model is adequate but has significant gaps in justification or effectiveness.	Model has many issues, is not validated, or is poorly justified.	Inadequate or no model development.	10%
Data Visualization	Exceptional, clear, and insightful visualizations. Demonstrates mastery.	Effective visualizations with minor issues in clarity or relevance.	Adequate visualizations but with some significant gaps or errors.	Poor choice of visualizations or significant errors in design.	Minimal or no relevant visualizations.	10%
Ethical Consideration	Demonstrates comprehensive understanding of ethical implications, identifies potential biases, and suggests actionable solutions.	Demonstrates good understanding of ethical implications and identifies most potential concerns.	Recognizes basic ethical concerns but lacks depth or fails to suggest solutions.	Minimal recognition of ethical concerns. Significant oversight or misunderstanding.	Ignores or fails to recognize any ethical implications.	15%
Report and Presentation	Report and presentation are clear, comprehensive, and effectively communicate insights.	Report and presentation are mostly clear and communicate most findings well.	Some clarity in report and presentation, but significant gaps in communication.	Report and presentation lack structure and clarity.	Report and presentation are incomplete or incoherent.	20%