**DATA SCIENCE FOUNDATIONS**

**BUAN 620**

**CATALOG DESCRIPTION**

This course provides an in-depth understanding of data, data technologies, and techniques essential to effective and ethical data analytics. The course also includes fundamental concepts and skills to effectively manage, clean, integrate, pre-process, and transform data for analytics using industry-standard tools such as Python or R.

*Prerequisite*: BUAN 600

**COURSE LEARNING OUTCOMES**

Upon successful completion of this course, students will be able to:

1. Describe the difference between the popular and technical understanding of data
2. Recognize different types of data attributes
3. Use industry-standard tools to read, manipulate and analyze data
4. Recognize analytics opportunities where data analytics could lead to more effective decision making
5. Discuss and demonstrate the role of data management systems and technologies for effective analytics
6. Connect to databases directly or through APIs
7. Perform data cleaning, integration, transformation
8. Recognize and handle outliers and missing values
9. Choose and implement the appropriate data transformation
10. Recognize ethical data modeling concerns, discuss their multiple positive and negative aspects
11. Recognize assumptions behind datasets and data models, and make decisions that are ethically and societally sound.

**COURSE OBJECTIVES**

* Distinguishing between popular and technical understanding of data, and learning about different types of data attributes
* Learning a programming tool that will help the analytic process
* Discussing the goals of data analytics, to empower effective and ethical data preprocessing
* Discussing the role of databases, and how data can be collected from them for analytics
* Discussing the importance of data cleaning, its many aspects via real examples
* Discussing the challenges of data integration and performing it in some real example
* Discussing the role of data transformation, and learning the most popular transformation techniques
* Discussing the ethical considerations of data modeling and the role of assumptions in the ethical data-driven decision making

**LEARNING MATERIALS AND RESOURCES**

**REQUIRED Text:**

Jafari, R (2022). Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics. Packt. ISBN- 978-1801072137

Amazon link to purchase a hard copy: <https://www.amazon.com/Hands-Data-Preprocessing-Python-effectively/dp/1801072132>

Publisher link to purchase e-book: <https://www.packtpub.com/product/hands-on-data-preprocessing-in-python/9781801072137>

**OPTIONAL TEXT:**

Mattan Griffel, Daniel Guetta (2022). Python for MBAs. Columbia Business School Publishing ISBN-13: 978-0231193931

Marc Wintjen (2020). Practical Data Analysis Using Jupyter Notebook: Learn how to speak the language of data by extracting useful and actionable insights using Python.

Packt. ISBN- 9781838826031

**REQUIRED SOFTWARE:**

Anaconda Navigator (Free installation from <https://docs.anaconda.com/anaconda/> )

**COURSE WEBSITE:**

The course materials including lecture notes, the syllabus, and teaching plans and goals will be accessible through Moodle (found via <https://learn.redlands.edu>). Students should regularly check the course website for announcements, etc. Students are also responsible for checking their redlands.edu email account or having this account forwarded to their preferred email account.

**PROGRAMING PREREQUISITE:**

Basic computer programing understanding such as variables, conditionals, and loops are required. If you have passed any programming course in any programming environment such as Python, Java, C, etc. you are set to start learning in this course. If you will be new to computer programming or it has been a while since your last programming exposure, you may check the Course Website under PROGRAMMING PREREQUISITE for resources to help you brush up on your basic programming skills.

**ASSIGNMENTS**

**Course Project**

To do this project successfully, you need computer data manipulation skills and data literacy. You will acquire these skills during the course. The open-source programming environment, Python, and its open-source modules, NumPy, Pandas, and Matplotlib, will be covered during this course.

For your course project, you need data. You may use one of the 10 suggested data sources in chapter 18 of the textbook. Each suggested dataset comes with a brief explanation of the data collection and guidance on the decision-making environment and how the analysis of the data can be beneficial.

You are not required to use the suggested data sources. You could find your own data. Good resources to find possible data sources are listed below. Students who wish to work on a dataset of their own choosing are highly encouraged to consult the instructor so the usability of the dataset for this project is assessed from the beginning and avoid too-late disappointments.

* <https://www.kaggle.com/datasets>
* <https://archive.ics.uci.edu/ml/datasets.php>
* <https://www.data.gov/>
* <https://aws.amazon.com/data-exchange/>

The project will be submitted in 4 phases. Each phase submission may have a redo option, meaning you can improve your submission and your grade by incorporating feedback from the instructor. Read the description of each phase carefully as each phase must include specific items.

For a sample of project delivery, you may read chapters 15, 16, and 17 of the textbook. Each chapter, with minor adjustments, could be project delivery.

**Phase 1 –** **Decision-making problem:** You submit a brief explanation of the decision-making environment. The explanation should clearly specify the decision-maker and include the major goals, the general objectives, and the current challenges of the decision-making environment. Also, you will formulate four questions whose answers can help with fulfilling the goals, achieving objectives, and overcoming the challenges. These questions should be in such a way that the data source you have access can help you answer them.

Submission should include one paragraph (500-word limit), four questions, and one image. The image can be created or straight from the internet to help you present the decision-making environment better. Explaining the decision-making environment by using the combination of words and images is much more effective. Use the image to help your audience better imagine the decision-making environment. Examples of such images can be seen in chapter 18 of the textbook.

**Phase 2 – Data:** This phase has a list of deliverables:

1. An introduction to the data that must include but is not limited to 1) how was the data collected, 2) why was the data collected, 3) how did you find the data. If you have more than one source of data, both should be included.
2. If applicable, report the data cleaning and integration steps you took to prepare the data for the next steps. If you believe data cleaning and integration is not necessary, justify.
3. Introduction of the data object
4. A table that introduces all attributes. Each row of the table presents one attribute. The table must include a description of the attributes if their names are not intuitive. Also, the table includes summarizing information such as the appropriate central tendency, the presence of missing values, the range of the attribute (max and min). In the case of categorical attributes, different levels of the attributes and their percentage should also be included.
5. To get to know the data better, for each continuous attribute, include the histogram or the boxplot of the attribute
6. To get to know the data better, for each categorical attribute, include the bar chart of the attribute

Submission should include one paragraph for Ⅰ (200-word limit), a sentence for Ⅲ, a table for Ⅳ, as many appropriate visuals as needed for Ⅴ and Ⅵ.

**Phase 3 – Data Preprocessing:** This phase has a list of deliverables:

1. Report on Missing values: Does the data have missing values? What types of missing values? Do the appropriate analysis to find the types of missing values in your data. What strategy is best to rectify each missing value? What is your rationale?
2. Report on extreme points:Find the extreme points. Analyze these extreme points and report if these extreme points are outliers or mistaken input? What should be done to each extreme? What is your rationale?

Submission should include the appropriate visualizations and analysis as needed.

**Phase 4 – Analysis and Conclusion:** This phase includes two parts. First, for each question in phase 1 create one or two visualizations (table included) that answer the question. Explain what each visualization presents and the conclusion you draw from it. Furthermore, now that you know the answers to the three questions put forth a general recommendation that can help the decision-making environment.

Second, provide the ethical guidelines that the decision-makers must know about when using the results of your analysis. What are some critical assumptions that have shaped the results? How and in what situations do the benefits of the models outweigh the costs of the assumptions? What are the roles of noise and bias in the collected data and how that can change the applicability of the results of the analysis?

**Project Deadline:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Phase 1** | **Phase 2** | **Phase 3** | **Phase 4** |
| **Hard Deadline** | Week 5 | Week 6 | Week 7 | Week 8 |

**Programming Assignments**

There will be seven programming assignments that help to solidify your learning. At the end of each session, an assignment will be given and they will be due next session.

**Other Assignments**

There will be other non-programming assignments such as mandatory discussion participation, or other activity that are assigned across the course.

Grades will be assigned according to the following **Number/Letter Grading Relationship Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| 4.0 A | 95% - 100% | 2.0 C | 73% - 76.99% |
| 3.7 A- | 90% - 94.99% | 1.7 C- | 70% - 72.99% |
| 3.3 B+ | 87% - 89.99% | 1.3 D+ | 67% - 69.99% |
| 3.0 B | 83% - 86.99% | 1.0 D | 63% - 66.99% |
| 2.7 B- | 80% - 82.99% | 0.7 D- | 60% - 62.99% |
| 2.3 C+ | 77% - 79.99% | 0.0 F | below 60% |

**Course Assessment and Grading**

|  |  |  |  |
| --- | --- | --- | --- |
| **ASSIGNMENTS** | **Linkage of Assignment to Course Learning Outcome(s) #** | **WEIGHT** | **SESSION DUE** |
| Course Project – Phase 1 | 4 | 10 | 5 |
| Course Project – Phase 2 | 2, 3, 6, 7 | 10 | 6 |
| Course Project – Phase 3 | 3, 7, 8, 9 | 10 | 7 |
| Course Project – phase 4 | 4, 10, 11 | 10 | 8 |
| Programing Assignment #1 | 3 | 5 | 1 |
| Programing Assignment #2 | 6 | 5 | 2 |
| Programing Assignment #3 | 3 | 5 | 3 |
| Programing Assignment #4 | 7, 8 | 5 | 4 |
| Programing Assignment #5 | 7 | 5 | 5 |
| Programing Assignment #6 | 7 | 5 | 6 |
| Programing Assignment #7 | 9 | 5 | 7 |
| Ethical Discussion | 10, 11 | 5 | 8 |
| Final Exam | 1, 2, 4, 5, 8, 9, 10, 11 | 15 | 8 |
| Other Assignments | 1, 2, 3, 4, 5, 6, 7, 8, 9 | 5 | - |
| **TOTAL** |  | **100%** |  |

**COURSE SCHEDULE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Required Reading** | **Workshop Topics & Activities** | **CLO** | **Deliverables Due** |
| **1** | Chapters 1&2 | 1. Class Orientation 2. Module 1 Technical foundations | 3 | * Programing Assignment #1 * Numpy Drag the Words Activity (Chapter1) * Pandas Drag the Words Activity (Chapter 1) * Matplotlib Drag the Words Activity (Chapter 2) |
| **2** | Chapters 3&4 | 1. Module 2 Technological and conceptual foundations | 1, 2, 5, 6 | * Programing Assignment #2 * DATA Drag the Words Activity (Chapter 3) * Databases Drag the Words Activity (Chapter 4) |
| **3** | Chapters 5-8 | 1. Module 3 Data Analytic Goals | 4 | * Programing Assignment #3 * Data Visualization Drag the Words Activity (Chapter 5) * Prediction Drag the Words Activity (Chapter 6) * Classification Drag the Words Activity (Chapter 7) * Clustering Drag the Words Activity (Chapter 8) |
| **4** | Chapters 9-11 | 1. Module 4 Data Cleaning | 7,8 | * Programing Assignment #4 * Data Cleaning Level 1 Drag the Words Activity (Chapter 9) * Data Cleaning Level 2 Drag the Words Activity (Chapter 10) * Data Cleaning Level 3 Drag the Words Activity (Chapter 11) |
| **5** | Chapter 12 | 1. Module 5 Data Integration | 7 | * Course Project Phase 1 * Programing Assignment #5 * Data Integration Drag the Words Activity (Chapter 12) |
| **6** | Chapter 13 | 1. Modules 6 Data Reduction | 7 | * Course Project Phase 2 * Programing Assignment #6 * Data Reduction Drag the Words Activity (Chapter 13) |
| **7** | Chapter 14 | 1. Module 7 Data Transformation | 9 | * Course Project Phase 3 * Programing Assignment #7 * Data Transformation Drag the Words Activity (Chapter 14) |
| **8** | -- | 1. Module 8 Ethical Foundations of Data Science | 10,11 | * Final Exam * Course Project Phase 4 * Ethical Discussion Assignment |