UMass | Dartmouth
UNIVERSITY STUDIES

Master Syllabus - *Annotated Template*
Course: *Introduction to Data Science, DSC 101*
Cluster Requirement: *Foundation for Learning through Engagement (1E)*

This University Studies Master Syllabus serves as a guide and standard for all instructors teaching an approved course in the University Studies program. Individual instructors have full academic freedom in teaching their courses, but as a condition of course approval, agree to focus on the outcomes listed below, to cover the identified material, to use these or comparable assignments as part of the course work, and to make available the agreed-upon artifacts for assessment of learning outcomes.

Course Overview:

This course will survey foundational topics in data science. Students will learn a broad range of data science skills applicable across different domains, including social sciences, finance, crime and justice, social networks, and engineering. Students will develop statistical and computational thinking skills, and they will apply these skills to real-world datasets. Specific topics include applied data problems, statistical software, data frames, descriptive statistics, natural language processing, data storage, data merging, linear regression, and data mining. The core skills developed in this course lay a foundation for more advanced coursework in data management, visualization, exploratory data analysis, and machine learning.

University Studies Course Rationale:

The design of this course centers around a concept called the data science life cycle. That is, we view tasks or steps in the practice of data science as forming a process, consisting of states that indicate how it comes into life, how different tasks in data science depend on or interact with others until the birth of a data product or a conclusion. Naturally, different pieces of the data science life cycle then form individual parts of the course. Details of each piece are filled up by concepts, techniques, or skills that are popular in industry. Consequently, the design of our course is both “principled” and practical. A significant feature of our course philosophy is that, in line with activity theory, the course is based on the use of tools to transform real data to answer strongly motivated questions related to the data.

The course starts with an introductory lecture of data science with two goals in mind. One is to give students a sense that data entails value, another that it is possible to make a difference, to influence outcomes, by leveraging values from the data. We introduce numerous interesting stories from a variety of fields, ranging from science, finance, metrology, sports, to Internet and ecommerce, on how insights can be obtained from the data through models and analytical tools. Of course, these stories also convey an idea to students of what constitutes data science, and how their activity, on raw data, with specific objectives, can transform that raw data to insightful outcomes through the use of appropriate tools. Then the data science life cycle is introduced, followed by various parts of the cycle, including asking interesting questions from data, data collection, exploratory data analysis, modeling, and confirmatory data analysis.

DSC 101 brings together basic skills in thinking, reading, writing and quantitative reasoning to obtain insights from a variety of datasets relevant to local, regional and global communities. This course forms the foundation of knowledge, skills, and the data science life cycle that will be developed throughout the data science major.

Learning Outcomes:

<u>Course Specific Learning Outcomes:</u>

After completing this course, students will be able to:

1. Understand and apply contemporary techniques for managing, mining, and analyzing data across multiple disciplines
2. Be able to use computation and computational thinking to gain new knowledge and to solve real-world problems
3. Communicate their ideas and findings in written, oral and visual form
4. Understand and apply the data science life cycle to a variety of datasets drawing from different academic disciplines and industries spanning local, regional, and global communities.
5. Define engaged learning in the context of data science and domain-specific communinities.

6. Explain how perspectives within one or more academic disciplines impact the interpretation of data.

University Studies Learning Outcomes:
After completing this course, students will be able to:
Cluster 1E:
1. Express the rationale for a broad education, as described in the UMD Commitment to Student Learning.
2. Define engaged learning in the context of their major, discipline or community.
3. Apply the concept of engaged learning to their personal goals.
4. Explain how perspectives within one or more academic disciplines impact the community.
5. Explain how issues in the community can be understood within an academic discipline.

Examples of Texts and/or Assigned Readings:

Since its creation in 2015, DSC 101 has been taught by the same instructor who has not required any textbook. Instead lecture notes are provided to the students. The following list provides the lecture note titles:

- Lecture notes for Introduction
- Lecture notes for Data Science life cycle
- Lecture notes for Sampling and data collection
  - Data Science is an essential part of the daily life of modern citizens. For example, data science skills can help one to better understand news or critically look at news that involves in-depth analysis based on data. As part of this unit, we consider the important of data collection (or sampling). We cover topics on various forms of biased data collection (which can lead to erroneous or unethical outcomes), how to analyze the data collection procedure to decide if any potential biases were introduced, and, of course, how to collect data in a statistically sound manner. These important skills are applicable to many disciplines that involve data collection (e.g., experiments or measurements etc) or analysis. Students are made aware of the importance of data literacy as an important foundation of their liberal arts education.
- Lecture notes for Descriptive and summary statistics
  - In modern society, data-driven decision-making has become more and more popular. Analysis or decision-making based on data inevitably involves uncertainty. In this unit we begin to consider how to make decisions and analysis in a principled way when our data and analysis includes uncertainty. DSC101 covers two approaches for different scenarios. The first is some exploratory data analysis. DSC101 dedicates several lectures on this, including data summary, visualization, data transformation, data engineering. We additionally have in-class practice and exam questions asking students to sketch visualization for some given data. The second is hypothesis testing, which provides a statistically sound framework for data-driven decision-making or confirmatory analysis. Hypothesis testing has also been widely adopted in many, if not all, subjects of study that involve experiments or data analysis. Lectures call attention to the connections between diverse subjects.
- Lecture notes for R programming
- Lecture notes for Data visualization
- Lecture notes for Data transformation and engineering
  - As part of this section, students are asked to find news or media coverage that reports on potentially flawed or misleading data analysis. We also ask students to collect data for their projects and to provide critical reasoning and sound mathematical justification that their collected data are suitable. DSC101 also has assignments on problems that are either direct application of hypothesis testing, or approach the problem within a hypothesis testing framework, maybe in a surprising way.

- Lecture notes for Visualization of multivariate data
- Lecture notes for Clustering (classical)
- Lecture notes for Modeling, Regression
    - In this unit, students are introduced to the idea that one can leverage data for value. Data is an essential resource for modern society and can serve many important societal purposes, and it is important for students to understand this principle at play in various contexts. Students are exposed to the potential use of the data they have, which could lead to new business models, entrepreneurship opportunities, or better public policy. By understanding the value of the data in different fields of inquire, students are better equipped to understand the need to protect, collect, and ethically use data like any other valuable resource. We also engage students in discussions about potential privacy issues involved in the data, and to guard against the misuse of data which can lead to inequities, bad policy decisions, or other detrimental effects. By viewing data as a resource that must be used in an ethical, scientifically appropriate way, students are better equipped to approach data collection and analysis issues that may arise in a broad range of coursework outside of their major.

Example Learning Activities and Assignments:

The following sample lab and project activities addresses Cluster 1E Outcomes 1-5.

For labs, students are required to produce a written report that combines both their analysis and insights. For Projects, students also give an oral presentation. These projects cover fundamental concepts to data science and the data lifecycle.

A grade is assigned to each student based on the quality and completeness of their work as well as the broader insights drawn from the data as well as application of these insights to different disciplines and communities. This is an important aspect of the projects and labs, which allow students to experience engaged learning in the context of data science and the data lifecyle. For example, students apply their data science skillset to answer important questions including identification of misinformation and understanding how data shape national health policy (lab 1), exploring census data that determines distribution of national resources (lab 2), and exploring trends in global terrorism which shape policy and may be misrepresented in the media (project 2).

For each lab assignment, students are asked to reflect on the benefits of a liberal education in the context of the analysis. Data science labs have been designed to promote broad knowledge (drawing from diverse areas including epidemiology, demography, crime and justice, genetics, global terrorism, and retail) and to think critically. Students will read the University's statement on Commitment to Student Learning (https://www.umassd.edu/universitystudies/umassdcommitment/) and are asked to write about this statement in the context of data science and, specifically, the questions, data collection methods, and insights that arise in each lab. The precise nature of these questions is tailored to each lab, and we give some examples below. Some of the broad themes focus around what are the values and assumptions underlying measurement and data collection procedure? How can faulty data collection lead to incorrect conclusions? How can we gain meaningful insight from data to yield value for decision making? How should we interpret our model's output in a broader societal context?

1. ***Lab***: From recent media or online sources, students are asked to find 2 examples of bad practice in sampling or data collection that are related to Covid-19 (any examples not related to Covid-19 do NOT count). Tell which type of bad practice that is for each of the two examples, and what are the potential statistical issues. Students are also required to read the online articles about the estimation of the true number of infected Covid-19 cases. The claims is that this number is at least 6 times as much as the reported. Discuss how you think of this analysis and how you might do this better. Students are asked to provide 2 examples of how public health policy is guided by Covid-19 data or statistical models. Students are then asked to consider how the examples of faulty data collection can lead to poor public health policies, and how this could impact different groups of people and local economies.

2. ***Lab***: Students are asked to find an example of data visualization from a recent news/article. Please provide a short description of the data, a screen shot of the figure as well as your comments on the visualization. Comments should address how visualization is being used to convey meaning from the data. What point is the visualization trying to make, and does it try to persuade the reader to some preferred point of view? What impact might the visualization have on a diverse group of readers?

Students then create their own data visualizations based on the following datasets:

Find US population by states according to census 2000 and census 2010, respectively.
2) Find locations of US states in terms of latitude and longitude.
3) Draw bubble plot of US population by states such that the size of the bubble is proportional to the population
Mark the states name by their two-letter abbreviations , for example,
Massachusetts by MA. Do this separately for year 2000 and 2010.
4) Calculate the ratio of changes in populations by states. Draw heat map of the ratio by states.

3. ***Lab***: Clustering of the US Arrest data

1. Use the following R command to obtain the US Arrests data

>data(USArrests);

2. Try K-means clustering on the US Arrests data with different number, k, of clusters.
a) Describe your findings at different k
b) In your opinion, which k works the best? Please give your explanation.
c) Plot your clustering results at the `best' k (you can choose to plot the data on the two variables that give the best visualization, or use the top 2 principal components if you like).

3. Try agglomerative clustering on the US Arrests data.
a) Plot the dendrogram
b) Tell what is the optimum height to cut the dendrogram for clustering. Does this agree with your result on 2 b)?

4. Try divisive clustering on the US Arrests data.
a) Plot the dendrogram
b) Tell what is the optimum height to cut the dendrogram for clustering. Does this agree with your result on 3 b)?

Students are then asked to interpret their results using different clustering algorithms. Imagine you are asked to critique a proposed public policy bill to help reduce crime. The bill provides more funding for high-crime areas. How could your clustering plots be used to assess the effectiveness of the bill? Do different clustering plots give different answers? If so, how can you best use your analysis to provide a clear, unbiased opinion of the new bill?

4. ***Lab***: Regression on Pearson's father-son data

b) Produce a scatter plot of father's height (x axis) Vs son's height (y axis). Example R command:
c) Add the regression line on the scatter plot
d) Report the linear regression output (including R^2 etc)

Students are asked to read online resources such as (https://statisticsbyjim.com/regression/interpret-r-squared-regression/) to consider situations where R^2 value may be misleading. Are the R^2 values from this analysis to be trusted? Why or why not? Based on your refection, do you trust your linear regression model for this dataset?

5. ***Project***: Car prices by year in current US market

For your favorite car model (e.g., Ford Taurus), find the average prices for cars of this model that were manufactured in each year from 1997 to 2015 (in US). You can do this from some web sites such as cars.com, truecar.com etc. When you process the data, try to store other information about the car, such as mileage, condition etc; you may need to use this data for project 1. Or, you can simply use the data in the given example here. Or, you can use a dataset from the UC Irvine Machine Learning Repository to carry out a linear regression and answer the similar questions. The following are requirements:

a) A detailed description on how you obtain the data (be cautious on potential bias in data collection).
b) For each year, you need to find the price of at least 100 cars, and then calculate the average
(click here for example on how to extract the car price from a messy text file)
c) Produce a year-price scatter plot (click here for an example)
d) Tell during which years the dip in prices slows down. If you want to buy a used car or if you have a new car, when would you buy or sell it? Why?
e) Carry out linear regression on average prices Vs year

Produce another scatter plot and add the regression line
Report output of the linear regression

g) Include the data (only the average prices and the years) as part of your submission

6. ***Project***: Global terrorism data

This is a dataset from kaggle.com, which consists of more than 150000 terrorist attacks during 1970-2015.

a) A detailed description on the dataset.
Define major attacks are those involving causalities more than 10, 3-10 as small attacks and minor otherwise. For each of minor, small, ad major attacks, complete b-d)

b) Produce a scatter plot of year Vs number of attacks for major attacks and minor attacks, respectively.

c) Tell if there were years when there are changes in the trend of #attacks Vs Year. By using online resources, please offer an explanation for these changes.

d) Carry out linear regression on #attacks Vs year
Produce another scatter plot and add the regression line
Report output of the linear regression.
Students are asked to read online resources such as (https://statisticsbyjim.com/regression/interpret-r-squared-regression/) to consider situations where R^2 value may be misleading. Are the R^2 values from this analysis to be trusted? Why or why not? Based on your reflection, do you trust your linear regression model for this dataset? Are you able to use your analysis to build a case for terrorist attacks increasing, decreasing, or both? That is, can the analysis be manipulated such that the data will support the conclusion that attacks are either decreasing or increasing? As a data-literate citizen, what steps can you offer to other students to look for such data manipulations?

7. ***Additional projects***: Projects change over the years. In the past we've given other projects. For example, one was about a large-scale study in untangling the relationship among smoking, low birthweight, and infant mortality. Another is on how an e-commerce web site may use historical transaction records to build an item recommendation engine.

Outcome Map:

Univ St Learning Outcome	Teaching and Learning Activities	Student Work Products
1 Express the rationale for a broad education, as described in the UMD Commitment to Student Learning.	Readings, lectures and Labs, covering the concept or idea of data science, the data science life cycle, and how to apply data science to real world problems and the development of new business models, as well as basic skills regarding data collection, data visualization, exploratory data analysis, and R programming etc	Assigned Labs and course project
2 Define engaged learning in the context of their major, discipline or community.	Readings, lectures, assignment related to articles from news media or current major events (such as COVID-19 pandemic), in-class discussions on data science ideas on some real-world problems, lectures on how data science may be applied to understand the world through the lens of data or to make impacts to the society	Assigned Labs and readings
3 Apply the concept of engaged learning to their personal goals.	Lectures, covering the connections to other majors such as computer science, mathematics and business, career opportunities of data science and the impact of data scientists to digital economy and business, case studies of data science in the real-world	Course project with topics from the real world
4 Explain how perspectives within one or more academic	Lectures, covering real-world examples, convey the idea that data science knows no boundary, including to economically less developed areas and to small businesses. In-class discussion on	In-class discussion writeup and presentations

disciplines impact the community.	possible applications that can be built from a collection of data, e.g., auto traffic data, can potentially help improve urban planning, road condition alert, or best time for shopping etc.	
5 Explain how issues in the community can be understood within an academic discipline.	Readings and lectures on the discipline of data science, analysis or visualization of datasets related to issues in the community help students understand such issues from the data science perspective and gain insights from the data	Assigned labs and projects

Sample Course Outline:

The following outlines, in chronological order, the specific topics covered. Note that throughout the semester students also take 2 exams, 5 quizzes, 5 in-class labs, and 4 in-class hands-on practice/tutorial sessions.

- Introduction to data science (1 lecture)
- The Data Science life cycle (1 lecture)
- Sampling and data collection (2 lectures)
- Descriptive and summary statistics (1 lecture)
- R programming (5 lectures)
- Data visualization (4 lectures)
- Data transformation and engineering (1 lecture)
- Visualization of multivariate data (2 lectures)
- Clustering (2 lectures)
- Modeling, Regression (2 lectures)
- Hypothesis testing (3 lectures)