Introduction to Data Science STT 180-001 **Spring 2021**

Instructor	Dola Pathak dpathak@msu.edu Office hours (via zoom https://msu.zoom.us/j/93007006215 Password: **SS21DOLAOH)** MW 3:00 -- 4:00pm (Other times by appointment).
Teaching Assistant	Sarah Manski manskisa@msu.edu Office Hour: TBA via zoom https://msu.zoom.us/j/98550135737 & Password: **STT180HELP**
Learning Assistants	Andrew Wilt (wiltand1@msu.edu) Jeremy Austin (austi120@msu.edu) Helproom Hours: TBA via zoom https://msu.zoom.us/j/98550135737 & Password: **STT180HELP**
Course Meetings	Mondays and Wednesdays 10:20 AM-12:10 PM via Zoom https://msu.zoom.us/j/97604487557 Password: **CLASSROOM1** *Two meetings per week, on MW are synchronous during the scheduled class time in Michigan (US Eastern Time Zone). Most of the work in this class is done in groups during class time. So it is highly recommended that students attend the classes and work with their groups. However, accommodations will be made for students who might not be able to attend synchronously because of time zone issues. Contact your instructor immediately if that is the case for you.*
Course Description	Pervasiveness and utility of data in modern society. Obtaining and managing data. Summarizing and visualizing data. Ethical issues in data science. Communication with data. Fundamentals of probability and statistics.
Course Mission	Step into the world of Data Science using R: This course will introduce you to the world of analytics using exploratory and statistical methods and help you to reason, critically think, analyze, interpret, and present information extracted from data!
Course Objectives	In this course you will learn about data ethics, how to manipulate data objects, produce advanced graphics, tidy and wrangle data, generate reproducible statistical reports using R markdown, learn about the fundamentals of probability and statistics, and rudimentary machine learning techniques.
Course platform and management	This course will be delivered online through **D2L**. You will need your MSU NetID to login to the course from the D2L homepage (http://d2l.msu.edu).

	In D2L, you will access the slides and videos, course materials, and additional resources. For instant communication and group works, this course will use **MS Teams**. The link to download MS Teams for mobile and desktop can be found in https://www.microsoft.com/en-us/microsoft-365/microsoft-teams/download-app
Prerequisites	A course in calculus (either MTH 124, MTH 132, MTH 152H, or LB 118) is required (can be taken concurrently)
Course structure and assessment components.	Please refer to the supplemental file in the "Getting started and navigating the course" folder.
Required Texts	• **R Programming for Data Sciences by Melfi, Finley, and Doser** (pdf in D2L (https://www.stt.msu.edu/~melfi/STT180Text) • **Naked Statistics: Stripping the Dread from the Data by Charles Wheelan** (available in the bookstore and at https://www.amazon.com/Naked-Statistics-Stripping-Dread-Data-ebook/dp/B007Q6XLF2). • **OpenIntro** (pdf in D2L).
Supplemental Texts	R for Data Science (https://r4ds.had.co.nz/index.html) R Cookbook (https://rc2e.com) Both are available in the library, online, and as inexpensive printed versions from Amazon. Neither is required
Required Technologies	**D2L, R, RStudio, and Teams**. Please have R/RStudio and Teams installed before the class starts. Instructions to install R/Rstudio is course content "Getting Started" folder. Please check availability in your current location. You will need a laptop to install R/Rstudio.
Technology Help	Help with D2L: **https://help.d2l.msu.edu**
Internet and Connectivity	The course requires access to **"high speed"** internet. Since the class is synchronous, and class work, exams will be done synchronously, it is very important to have **reliable internet and webcam access**. If ever, students have accessibility issues during synchronous class or exam sessions, they should immediately inform the instructional team and provide proof of the issue.
Grading	Participation (ICA and pop quizzes), 20% Exams (3 in total), 23% Homework, 27% Project, 30%

Grade Cutoffs Let X be a student's final grade computed as a percentage. A student's final grade on the 4.0 scale will be determined as follows:

$\mathbf{4.0}$	$\mathbf{90 \leq X \leq 100}$	$\mathbf{2.0}$	$\mathbf{70 \leq X < 75}$
$\mathbf{3.5}$	$\mathbf{85 \leq X < 90}$	$\mathbf{1.5}$	$\mathbf{65 \leq X < 70}$
$\mathbf{3.0}$	$\mathbf{80 \leq X < 85}$	$\mathbf{1.0}$	$\mathbf{60 \leq X < 65}$
$\mathbf{2.5}$	$\mathbf{75 \leq X < 80}$	$\mathbf{0}$	$\mathbf{X < 60}$

RCPD To arrange for accommodation a student should contact the Resource Center for People with Disabilities at https://www.rcpd.msu.edu or 517-353-9642

Important Dates

01/11/2021	Class begins
01/25/2021	Last day to change to or from CR/NC or Visitor (5:00pm) Open adds end (8:00pm)
03/10/2021	Last day to drop with no grade reported (8:00pm)
04/21/2021	Class ends
04/29/2021	Final Exam (7:45 -9:45AM)

Late Work Policy: Late work will not be accepted. However, under exceptional conditions please inform the instructional team before the due date and time so that accommodation can be made accordingly (please provide documented proof for late work). There might be a penalty for late submission.

Attendance policy Attendance is mandatory for synchronous online classes like STT180. If a student is unable to attend any sessions, owing to health or other reasons, it is important for you to communicate with the instructional team at the earliest to avoid points being deducted for non-participation.

Student continuity In a situation if you are unable to engage in course content for a prolonged period it is critically important for you to communicate with your instructor promptly. We anticipate that some students may encounter interruptions to their studies for any number of reasons (e.g., illness, need to provide medical or childcare, sustained loss of internet, etc.) and have plans in place accordingly. To provide built-in flexibility to students as they progress through their coursework this semester, the lowest score of the in-class assignments and 3-2-1 assignments will be dropped at the end of the semester before grades are tabulated.

In extreme cases where a student cannot reliably progress through course content for more than one week, they should inform the instructional team immediately, and If the student can provide documentation of the obstacle they face, we intend to work hard to accommodate their situation with empathy.

Instructor continuity We have an instructional team for STT 180 and we will have pre-recorded videos for the class. Thus, we do not anticipate any issues arising if the instructor is absent for an extended period.

Learning objectives and tentative schedule for STT 180:

Week	Module		Date	Assigned Reading
1		• Introduction • Install R/Rstudio • Pre-semester survey • Navigate the class in D2L. • Enroll into the MS teams class.	Jan 11-13	Installation guide
2	1	• Understand the various component of R/Rstudio (wd, working environment, console, source pane), how the use help for arguments and functions.	January 20	An Introduction to R: chapter 1.1-1.7

3		• Understand the basic know how of the R programming language
		  ○ Function(arg, arg, opt arg)
		  ○ Apply basic functions, mean(), sd(), which.min(), etc. is.NA()
		  ○ Store values as objects (using <-, =)
			January 25	RPDS: chapter 1, 2.4 – 2.6

		• Use concepts of data ownership, possession, and stewardship to interpret a data analysis scenario. Define intellectual property in the context of data science. (Assessment 2-3 scenarios of concepts) • Given a context, identify a possible ethical issue involved in data analysis. Read the article (cc can predict things that is considered private and decide whether its ethical to do so or not).	January 27	Slides and video on concepts of ethics https://www.thedailybeast.com/how-visa-predicts-divorce
		Homework 1 assigned	January 30	
4		• Rmarkdown. ○ Decide if it's better to use .rmd files or .r files.	February 1	NS: chapter 1, RPDS: chapter 3

	2	• Understand data structures in R
		  ○ Define the coercion hierarchy and, given a list of data formats, be able to rank the items in the list according to the hierarchy.
		  ○ Understand vector recycling and vectorization property in R
		  ○ Extract information from data frames and lists. Be able to extract info from dfs & lists in more than one way.
		• Load into RStudio .csv, .rda, and files from URLs
			February 3	NS: chapter 2, RPDS: 4.1 - 4.8

5		• Use the Base R functions
		  ○ Apply names, rownames, within, apply, etc. functions to data structures
		  ○ extract information
		  ○ do transformations (subsetting).
			February 8	

		• Use tidyverse ○ Familiarize with tidyr and its functions	February 10	RPDS: 6.1 - 6.4
6		• Use tidyverse ○ Familiarize with dplyr packages and its functions	February 13	RPDS: 6.5
		Exam 1 covering module 1 and 2 synchronous online	February 15	
		Homework 1 due	February 18	

7	3	• Conduct Exploratory Data analysis to identify and summarize the structure of and variables and the relationships between them. Do this by:
		  ○ Identifying the appropriate descriptive to summarize a variable's distribution
		  ○ Computing & interpreting summary statistics
		  ○ Creating graphs that display contrasts and relationships
		Create graphical representations of summary statistic
			February 22	NS: chapter 3,
				RPDS: 5.1 - 5.2

		• (contd)Conduct Exploratory Data analysis to identify and summarize the structure of and variables and the relationships between them. Do this by:
		  ○ Identifying the appropriate descriptive to summarize a variable's distribution
		  ○ Computing & interpreting summary statistics
		  ○ Creating graphs that display contrasts and relationships
		  ○ Create graphical representations of summary statistics.
			February 24	NS: chapter 4,
				RPDS: 5.3 - 5.6

		Homework 2 assigned	February 27	
8		• Given a context or research question, identify the most appropriate graph in response and produce it with ggplot() • Identify ethical issues involved in data visualization. If students are asked to create two visualization of a dataset, decide which one is a more honest representation of the data structure	March 1	
9	4	• SWBAT write simple functions • SWBAT analyze a function in terms of its efficiency and readability	March 8	
		• Use conditional statements. • Write for, while, repeat loops Explain factors influencing efficiency of loops by citing advantage of R's vectorization property or using pre-existing functions.	March 10	

10		• Write for, while, repeat loops Explain factors influencing efficiency of loops by citing advantage of R's vectorization property or using pre-existing functions. • Review for Midterm 2	March 15	
		Exam 2 covering module 3 and 4 synchronous online	March 17	
		Homework 2 due	March 20	
11	5	• Understand simulations and some associated functions(sample (), replicate (), rnorm (), etc.) • Define the sampling distribution of a sample statistic as a model of the behavior of that statistic (in terms of center and dispersion) for repeated RS of the same size. ○ Identify where the distribution is centered ○ Describe what the variability of the distribution signifies. • Anticipate the characteristics of a sampling distribution using the CLT. Identify instances where the CLT does not help predict the characteristics of a sampling distribution.	March 22	NS: Chapter 7, OIS 4.1-4.4, 6.1
		• Describe the roles sampling distributions and CLT play in inference.	March 24	NS: Chapter 8 OIS 5.1
		Homework 3 assigned	March 27	
12		• Given a context, express a research question in terms of two competing claims about a model parameter.	March 29	NS: Chapter 9
		• Do hypothesis testing and build confidence intervals for one mean, median, and proportion.	March 31	NS: Chapter 10, OIS 4.4-4.5
13	6	• Evaluate whether a bivariate dataset can be represented by linear models. • Compute and extract simple linear regression models from R. • Use residual diagnostics to assess whether OLS conditions are met.	April 5	RPDS: Chapter 9
		• Formulate and interpret linear models with categorical predictors • Interpret multiple linear regression coefficients and model fit.	April 7	NS: Chapter 11, OIS 7.1-7.3
		Homework 3 due	April 10	
14		• Perform KNN analysis • Use Cross-validation to determine the best K in the KNN algorithm. • Describe K means learning algorithm.	April 12	NS: Chapter 12, 13, OIS 7.4, 8.1-8.3
		Exam 3 covering module 5 and 6 synchronous online	April 14	

		Presentations due	**April 17**	
15	**7**	**Learn some of the essentials data scientist tools:** ○ **Relational databases** ○ **SQL in R**	**April 19**	
		Presentations	**April 21**	
16		**Project write-up due**	**April 26**	
		Peer Review Due	**April 28**	
		Final Exam: Thursday, April 29 2020 7:45am - 9:45am (**Note: there is no final exam for this class**)	**Dec 18**	
Assessments		• **In class group assignments during each class** • **3-2-1 assignments every week** • **Homeworks** • **Exams** • **Group Project** • **There maybe pop-up quizzes during the class session from the preassigned readings/videos.**		

*RPDS: R Programming for Data Science, NS: Naked Statistics, OIS; Open-Intro Statistics.