



GSMDS 5001 (Fall 2019)

Foundations of Data Science with R

Saturdays, 1:00 to 4:00 pm

Instructor: **Jameson Watts, Ph.D.**

Email: jwatts@willamette.edu

Office Hours: After class

Course Description: This foundational course offers a full-spectrum introduction to data science and data science workflows, emphasizing data as a source of value creation in an organization. The R programming environment serves as the implementation vehicle in support of essential data science activities—data exploration and visualization, data wrangling, predictive modeling, model deployment, and communication. The R programming environment, along with Python, is among the most important tools in the data scientist’s toolbox. This course will feature tools and a style of programming inspired by the popular tidyverse ecosystem—ggplot2 for data visualization, dplyr and tidyr for data wrangling. Students will master elements of the data science workflow through a series of short R programming exercises reinforced by a full-spectrum, integrative final project. Presentation skills are an ever-present theme as students are challenged, through every stage of analysis, to communicate managerial relevance and value to an organization.

Course Format: This course employs various pedagogies, including formal presentations by the instructor, case discussions, simulations, and in-class activities—the approach used depends largely on the class material for a given week. Active participation is paramount to your success in this course. Students are expected to question, challenge, or clarify the material as it is being presented, and to discuss issues/questions raised by your colleagues and/or the instructor.

Required Materials:

1. Base R, [Install here](#)
2. RStudio 1.2, [Install here](#).
3. R for Data Science (R4DS), [Read here](#)
4. DataCamp Classroom (DCC), [Join here](#)
5. Assorted articles, tutorials and links to media (posted to WISE).

Course Learning Objectives:

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

1. Write basic R code, functions, and markdown for analysis and reporting
2. Implement an efficient data science workflow including the use of version control software to work on teams
3. Import and manipulate large datasets within RStudio
4. Select important variables, filter out key observations, create new variables, and compute summaries
5. Use data visualization, summarization, and correlation to explore relationships between variables and both ask and potentially answer interesting questions about data
6. Implement a linear model and interpret the results
7. Design and analyze a business experiment
8. Describe the reliability and impact of results from a linear model
9. Choose an appropriate visualization for a given dataset and separate visualizations by category
10. Communicate your results in writing and in person, in a format appropriate for consumption by non-data scientists.

Course Outline:

Class	Date	Class Topics	Reading and Assignments
1	09/07	Course Overview and Example Analysis	Read the syllabus; Install R; Install RStudio 1.2; Join the DCC; R4DS: Ch. 1, 2; Read this article
2	09/14	Intro to R, Markdown and the Tidyverse	DCC: Explore your Data; DCC: Introduction to RMarkdown; R4DS: Ch. 4
3	09/21	Data Manipulation	DCC: Tame your Data; DCC: Tidy your Data
4	09/28	Data Wrangling	DCC: Transform your Data; R4DS: Ch. 12, 13
5	10/05	Exploratory Data Analysis and Basic Visualization	DCC: Grouping and Summarization; DCC: Data Visualization; R4DS: Ch. 7
6	10/12	Review and Midterm I	Study!
7	10/19	Modeling I: Basic Regression	DCC: Introduction to Modeling; DCC: Modeling with Basic Regression; R4DS: Ch. 22, 23
8	10/26	Modeling II: Business Experiments	DCC: Mini case study in A/B Testing; Mini case study in A/B Testing Part 2
9	11/02	Modeling III: Multiple Regression, Confidence and Impact	DCC: Modeling with Multiple Regression; DCC: Model Assessment and Selection
10	11/09	Advanced Visualization	DCC: Custom ggplot2 themes; DCC: Creating a Custom and Unique Visualization
11	11/16	Shiny Apps and Reporting	DCC: Building Static Dashboards; DCC: Customizing your RMarkdown Report
12	11/23	Review and Midterm II	Study!
Fall Break			
13	12/07	R Programming and Teamwork	DCC: Basic Workflow; R4DS: Ch. 17, 18, 19, 20, 21
14	12/14	Final Presentations!	Final Report due at beginning of class

Summary:

Assignment	Percentage
Homework Assignments	25%
Midterm Exam I	25%
Midterm Exam II	25%
Final Presentation	25%
Total	100%

Grade Distribution:

≥ 95.00	A
90.00 - 94.99	A-
85.00 - 89.99	B+
80.00 - 84.99	B
75.00 - 79.99	B-
60.00 - 74.99	C
≤ 60.00	F

Assignments:

- **Homework Assignments (25%):** Homework assignments consist of completing the assigned chapters within the DataCamp Classroom. Each week there will be one or two chapters to complete. Each chapter should take around an hour to finish. Your scores on the embedded programming exercises will constitute your homework grade.
- **Midterm Exams (50%):** We will have two midterm exams worth 25% of your grade each. We will spend the first hour of class reviewing past material. The exam will then be administered during the second two hours of class. You can expect to be given a dataset and a series of questions to answer using the skills developed during the course. Exams are open everything (book, notes, internet), EXCEPT communication with others.
- **Final Report and Presentation (25%):** By the third week of class, you will need to find a dataset that you wish to analyze. This data must be freely available (not subject to copyright) and free from prior analysis (no cheating off of other people's code). It also needs to be distinct from any datasets used in your Python class. If you want to use data from a company that you work for, you must first show me that you have gotten permission to do so. There are many sources of data online; however, I suggest that you start by looking at the sites listed [here](#).

On the final day of class, you will give a 15 minute presentation of your analysis. This must incorporate all appropriate skills learned in class and be delivered such that a non-technical person can follow along. You will also turn in a pdf report and the RMarkdown file that generated it. You will be graded on the clarity of your exposition, the comprehensiveness of your analysis, and the quality of your code.

Course Policies:

- Name tents must be used for the first month of classes—I am a chronically absent-minded professor, and pretty bad with names. I promise an attempt to memorize everyone’s moniker, but name tents guarantee that my deficiency does not lead to continuous embarrassment.
- Collaboration is encouraged both during and outside of the classroom. Students may work together to prepare for exams; however, each student will be on their own when taking the exams in class.
- No late assignments will be accepted under any circumstance except in rare cases of personal or family emergency.
- Students with disabilities who require accommodation should notify me of the nature of accommodation you require in the first week of class. Additional support is available from the Willamette University Accessible Education Services Office (www.willamette.edu/dept/disability), telephone 503-370-6471.
- Students are responsible for all missed work, regardless of the reason for absence. It is also the absentee’s responsibility to get all missing notes or materials.
- Every student is expected at all times to abide by the Willamette University Atkinson Graduate School of Management Honor Code (<http://www.willamette.edu/mba/about/honorcode>).
- You must also abide by the Application to Academic Honesty as detailed in the current student handbook (<http://www.willamette.edu/mba/students/student-handbook>).