PA 434
DATA ANALYTICS

Tuesday, 3:30 - 6:15 PM - Blackboard Collaborate

Federica Fusi
ffusi@uic.edu
Office hours available upon request (just send me an email!)

TA - Alea Wilber
awilbu@uic.edu
Office hours TBD + availability upon request

COURSE OVERVIEW

R is an open source language and environment for statistical analysis, data science, data wrangling and modelling, and data visualization. Over the years, R has become one of the most used analytics tools in the private and public sector. Large companies such as Google and AirBnB are currently using R for data analysis and visualization. International organizations, such as the World Bank, are also recognizing the importance of R for economic analysis. The New York Times uses R for data visualization. Given the widespread diffusion and constant growth of R, it has become one of the most important tools for those aspiring to "work with data"

This course is designed to achieve two objectives (1) make students comfortable with using R for data wrangling, exploration, and visualization and (2) make students comfortable with managing, cleaning, and analyzing new datasets, especially when presented in a 'raw' format.

1. The course begins with developing a basic understanding of the R working environment, including arithmetic and logical operators, getting help using R and the R community, data structures, variables, and data types. Next, students will be introduced to Tidyverse and will cover dplr, readr, ggplot2, tibble, stringr, tidyr, and purrr. We will utilize these various packages to work on both datasets available in R and publicly available open data at the city, state, and federal level. Students will learn how to import data sets and transform and manipulate those datasets for various analytical purposes. Students will also learn how to write R scripts and build R markdown documents to share their code with others - this includes addressing issues related to reproducibility and writing appropriate methodological notes. Finally, students will learn how to create control structures, such as loops and conditional statements to traverse, sort, merge, and evaluate data.
2. By the end of this course, you will be data literate and you will be able to ask questions about public problems and answer your own questions using publicly available data and statistical analysis and data visualization tools. We will discuss common issues affecting raw datasets, from data structure to missing data, sampling…. You will be able to turn raw data into policy insights, information, and ultimately understanding and knowledge for public organizations.

LEARNING OBJECTIVES

- Be curious and confident with data
- Feel comfortable with R
- Understand limitations of your data
- Being autonomous in exploring new R packages and functions.
- Find, identify, open new data
- Find patterns in data with appropriate graphs and visualizations
- Pose questions related to public problems
- Provide insights to answer these questions using basics descriptive statistics
- Accurately interpret results from basic inferential statistical analyses
- Communicate the results of your analyses in accessible language to policy makers (including data visualization)

COURSE MATERIAL

Relevant materials and readings will be provided on Blackboard 2 to 3 weeks in advance before the class is scheduled.
If you are curious about the content of the class, we will mostly cover the content in these main books:

- **Art of Data Science:** https://leanpub.com/artofdatascience (theory book - pick the "book only" option to download it for free)
- **ModernDive:** https://moderndive.com/ (R book)
- **R for data science:** https://r4ds.had.co.nz/ (R book)

All students are required to install R and RStudio on their laptop. I have provided instructions via email and on Blackboard.

ADDITIONAL GOOD BOOKS

- R for Data Science https://r4ds.had.co.nz/index.html
- ModernDive https://moderndive.com/index.html
- Data Visualization: A Practical Guide https://socviz.co/index.html#preface
- R Graphics Cookbook https://r-graphics.org/
- Fundamental of Data Visualization https://clauswilke.com/dataviz/
- Ggplot2: Elegant graphs for data analysis https://ggplot2-book.org/introduction.html
- Basic R Guide for NSC statistics https://bookdown.org/dli/rguide/
- Hands-on programming with R https://rstudio-education.github.io/hopr/
- Advanced R https://adv-r.hadley.nz/

ONLINE LECTURE POLICY

For our meetings, please plan to have your camera on. I totally understand that interruptions might occur during the class (e.g., kids, pets, short breaks to grab some water or stretch your legs) and you are allowed to occasionally turn off your camera. But I found that face-to-face interactions, even if digital, help with attention and retention and make the class a bit more

lively. So, make yourself comfortable and be ready to join! You can use a background image if you want to.

I might also occasionally ask you to share your screen to comment on your coding, so make sure to close any tabs or documents that you don't want others to see.

GRADING & ASSIGNMENT

The course is designed to accompany you through learning R. It progressively builds on your capacity to expand upon what you have learned in class - so, don't expect that exercises will always just use what you learnt in class!

Note: this course is primarily designed for those who have no experience in R or programming nor have extensively worked with data before. However, I am expecting students to have different skill levels and will do my best to address topics that might be of interest to more advanced students.

All assignments are due at midnight of the indicated date

Class participation (10% - 15 points)
Due: ongoing

You are expected to attend the weekly sessions. If you cannot attend, please let me or Alea know in advance.

In class, you are expected to contribute to everyone's learning by:

- Asking relevant questions
- Sharing your ideas with the class
- Sharing mistakes and problems you encounter while coding: debugging each other's work is the best way to learn fast
- If you want to: briefly present relevant packages and functions to the class (contact me about this)

As part of your participation grade, I will ask you to prepare questions to ask to your assigned discussant group during final presentations.

12 Weekly labs (40% - 60 points, 5 each assignment)
Due: Every Monday before class

These are short weekly exercises to be completed at home to get familiar with the various topics. Given the current circumstances, I understand that it might be difficult to maintain a weekly commitment; because of this, you all have *two* free passes to submit assignments late (try not to wait too long) without penalty nor need to communicate with me.

We will try to provide some individual feedback but I encourage you to take notes of difficulties or issues that you encounter while doing the assignments and share them with the class the following week. We will address them as a group.

In-class activity (10% - 15 points)
Due: March, 8th (Monday)

At the end of the first part of the class, we will do an 'in-class activity' to familiarize ourselves with working with real data. We will start the activity in class but you might need to complete it later on.

Final project (40% - 60 points)
Due: May, 4th

For your final, you will work in small groups that will be assigned by me. With your group, you will choose a public problem of interest to you and find relevant data to answer a set of relevant questions. Your scope is to tell a story with your data and present the problem and possible solutions to relevant government stakeholders. Your final project will include three parts:

1. A preliminary 1-2 pages proposal where you are expected to discuss your topic, question(s), and possible data **(5% - 7.5 points - Due: March, 29th)**
2. A class presentation of 6-7 minutes where you will give an overview of your problem, data, and preliminary results. The scope of the presentation is to collect feedback and questions to improve your final work. **(10% - 15 points - Due: April 20th or 27th)**
3. A full report to be submitted by May 4th which will include: **(25% - 37.5 points - Due: May, 4th)**
    1. a R source file with appropriate comments and documentation to explain your code. The R file should be perfectly reproducible by me and your TA
    2. A final report in PDF format (you are expected to produce this with R markdown) containing both a descriptive analysis of data and appropriate data visualization.
    3. A methodology write-up (this can be an appendix to your report) detailing data source, data characteristics, and limitations.

Team work
All assignments are to be submitted individually. However, teamwork is encouraged especially to address coding doubts and questions. Peer learning is important. Take advantage of one another to form a cohesive class group and help each other on solving lab assignments. If you collaborate with others, please report their names on the top of your assignment. This helps me to assess common mistakes. Additionally, make sure that your assignment is individually developed - while you are encouraged to work together on coding, comments, text, ...should be developed by each student individually.

GRADING POLICY

150-135 pts	**A - Exceptional**
134-120 pts	**B - At expectations**

119-105 pts	**C - Below expectations**
< 104 pts	**D - Insufficient**

CLASS SCHEDULE

Important: I am expecting that this syllabus schedule might be modified as we go depending on class progress and interests. There are also a few "TBD" sessions - this is because I want to have some flexibility if we need to spend more time on certain topics or the class takes a new direction. Any modifications to the syllabus will be posted on Blackboard in advance and email notification will be distributed to course participants. Modifications won't affect the assignments nor the grading policy.

CLASS 1 - Introductions
Jan, 12th 2020

Introduction to the class
Student introductions
Introduction to RStudio interface.

CLASS 2 - Traditional R
Jan, 19th 2020

A brief introduction to the traditional "dollar sign" framework to work in R - objects, vectors, matrices, and dataframes.

CLASS 3 - Working with datasets: tidy data
Jan, 26th 2020

Structure of a dataset, types of variable, tibbles

CLASS 4 - dplyr
Feb, 2nd 2020

Introduction to dplyr, Part 1

CLASS 5 - dplyr & exploring dataset
Feb, 9th 2020

Public problems
Introduction to dplyr, Part 2

CLASS 6 - Data visualization with ggplot2, Part 1
Feb, 16th 2020

Review of visualization rules.
How to visualize different types of data.
Basics of ggplot2

CLASS 7 - Data visualization with ggplot2, Part 2
Feb, 23rd 2020

More work with ggplots
Image formats

Class 8 - Recap exercise (in class activity)
Mar, 2nd 2020

Data science "in practice": Asking questions and exploratory analysis

CLASS 9 - Reproducibility and communication
Mar, 9th 2020

Introduction to R markdown
Workflow and reproducibility: why it matters?
Discussion about in class exercise

CLASS 10 – Working with factors
Mar, 16th 2020

forcats package
Ordering levels in ggplots

March 23rd – Spring break

CLASS 11 - Programming
Mar, 30th 2020

Exploring loops and apply commands

Due: Part 1 of the final project

Class 12 - Working with strings and other data types
Apr, 6th 2020

Working with the stringr package
What is text analysis? How can you work with text data?

Class 13 - Twitter data and text analysis
Apr, 13th 2020

A quick introduction to APIs and packages to download data with R
Using rtweet to collect Twitter data

Visualizing Twitter data
Discussion: are social media data “good” data?

Class 14 – Tables
Apr, 20th 2020

Producing nice tables with R
Due: Final project workshop

CLASS 15 – Programming Part 2 (tentative)
Apr, 27th 2020

Writing functions
Due: Final project workshop

COURSE AND UNIVERSITY POLICIES

Academic Integrity: UIC policies on academic integrity will be strictly enforced in this class. Instances of academic misconduct by students will be handled pursuant to the Student Disciplinary Policy (here). Please contact me if you have any questions about these policies.

Disability Accommodations: The University of Illinois at Chicago is committed to maintaining a barrier-free environment so that students with disabilities can fully access programs, courses, services, and activities at UIC. Students with disabilities who require accommodations for access to and/or participation in this course are welcome, but must be registered with the Disability Resource Center (DRC). You may contact DRC at 312-413-2183 (v) or 773-649-4535 (VP/Relay) and consult the following: http://drc.uic.edu/guide-to-accommodations. Please talk with me about any accommodation you may need to succeed in this class.

Plagiarism: Borrowing any idea, theory, information, or facts that are not common knowledge without acknowledging the source is considered plagiarism and it is a very serious offense. The academic honor code applies under all conditions and instances of plagiarism will be dealt with following UIC academic integrity policies. PLAGIARIZED ASSIGNMENTS WILL RECEIVE ZERO POINTS.