DATA 115: Introduction to Data Analytics

Fall 2024

Course Logistics

Prefix/Number:	DATA 115.1 (WSU Global Campus)
Credit Hours:	3
Prerequisites:	None
Location:	Global Campus

Instructor Information

Instructor:	Gani Nurmukhametov
Office:	WSU Everett Room 404
Phone:	(425) 405-1659
Email:	gani.nurmukhametov@wsu.edu
Office Hours:	typically[1] M 11 am - 12 noon, F 9-10 am or by prior appointment in Zoom

Introductory Note

First of all, welcome to DATA 115! I hope I will get a chance to meet you in person sometime during your journey at Washington State University.

Secondly, sorry this is such a long document. Despite that, I hope you will make time to read it through at least once during the first week of the course as it contains a useful summary of the material we will be covering and will serve as a repository of important information and links.

Finally, definitely feel free to reach out if you have any questions or concerns about the course or material. I am always happy to chat about data, statistics, or academic life and would welcome opportunities to offer my perspective or simply serve as a sounding board.

Course Description

The course describes basic concepts, principles, and tools used in data analytics. More specifically, this course provides an introduction to the field of data analytics. We will draw on recent and relevant materials from statistics, mathematics, computer science, artificial intelligence, and machine learning, as well as many application domains.

As befits a rapidly developing, interdisciplinary subject, we will draw on recent and relevant materials from statistics, mathematics, and computer science, as well as many application domains. Motivated by natural questions that arise in simple data examples, we will cover many of the basic techniques for working with data including sourcing raw data, cleaning and processing, exploring and analyzing, and finally presenting conclusions. In order to provide a foundation for later courses in the major, we will also explore initial examples of many of the core topics that will be encountered. You will have plenty of opportunities to work with real data and various tools (primarily Python, but also some Excel and R).

In addition to familiarizing you with basic tools and methods, this course will provide a broad exposure to the diverse types of data analytics projects that are being conducted around the world. A key component of the course will be critically analyzing published data analytics works and discussing their strengths and shortcomings. Finally, as data-driven practices are becoming common in many career fields, we will focus on professional development topics such as presentation skills and examples of the ethical and legal issues that can arise in modern data analysis projects.

[1] I may change the time or even cancel office hours on a given day. Please follow my announcements on Canvas.

Remote Logistics

Given the nature of the Global Campus course, all course materials will be presented entirely online, mostly on the Canvas website, with some videos displayed over the Panopto system. Generally, each week will present you with pre-recorded videos covering some of the course content along with lecture slides, notes, and assignments.

Course Materials

There is no required textbook for this course, as we will be making use of open source and online materials. The following six textbooks below (first three are for Python, the last three are for R) are freely available either as .pdf files or online and will be referenced during lectures and assignments:

1. An Introduction to Statistical Learning with Applications in Python by James, Witten, Hastie, Tibshirani, and Taylor; hereinafter referred to as *ISLP*.
2. Python Data Science Handbook: Essential Tools for Working with Data by VanderPlas; hereinafter referred to as *PDSH*.
3. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter by McKinney, 3rd edition; hereinafter referred to as *PDA*.
4. An Introduction to Statistical Learning with Applications in R by James, Witten, Hastie, and Tibshirani, 2nd edition; hereinafter referred to as *ISLR*.
5. R for Data Science by Wickham and Grolemund, 2nd edition; hereinafter referred to as *RDS*.
6. ggplot2: Elegant Graphs for Data Analysis by Wickham; hereinafter referred to as *ggplot2*.

There are many more different books that can help when learning Python and Data Science. The resources listed below are not open source and hence are not required for the course assignments; however, you may find them useful for self-study. Many of these books are available for free trial period, so consider using that option before committing to a purchase and/or rental.

- A Hands-On Introduction to Data Science by Shah.
- Data Science from Scratch: First Principles with Python by Grus, 2nd edition.
- Doing Data Science: Straight Talk from the Frontline by O'Neil and Schutt.
- Murach's Python for Data Science by McCoy, 2nd edition.
- Learning Data Science: Data Wrangling, Exploration, Visualization, and Modeling with Python by Lau, Gonzales, and Nolan.

In addition to these textbooks, some shorter weekly readings, data sets, and programming scripts will be uploaded on the Canvas course website.

Software

No previous programming experience or knowledge of statistical software will be assumed.

At the start of the course, we will focus on manipulating and extracting details from data using Excel.

Later segments of the course will provide training and examples using Python. Python is an open source programming language with a robust ecosystem of packages for data analytics that has become the predominant programming environment in data analytics. We will start with the basics of Python 3 and build up familiarity with standard data science libraries including numpy and pandas.

At the end of the course, we will have a short introduction to R, which is another open-source programming widely used in data analytics. We will learn the standard data science libraries like dplyr, ggplot2, knitr, tibble, and tidyr.

Class Communication and Instructor Interaction

We will use the Canvas forums for course discussions. This is a great place to ask questions from your peers, as well as to get feedback on your ideas. Announcements and other official communications will be posted on the Canvas website as well as sent to your official WSU email accounts. You should check these messages regularly to stay informed about upcoming due dates and updates to the syllabus.

I am accessible by email at gani.nurmukhametov@wsu.edu. I teach multiple courses, so please, as a courtesy to me, include "DATA 115 (Fall 2024)" in the subject line for any messages concerning this course. I will most likely read your email as soon as I receive it, but it still may take me a little while to respond. I will try to get back to you as quickly as I can but this does mean that queries sent immediately before a deadline may not receive substantive responses in time to be directly helpful, so please plan ahead ☺

I strongly encourage you to utilize the resource that is freely available to you, that is my office hours! This will allow you to get immediate feedback from me. I hold my office hours in Zoom. Following the Zoom meeting link, you will appear in the waiting room first, and if I am not already talking to another student, then I will accept you right away. If you cannot attend my office hours, do not hesitate to send me an email and we can schedule a short appointment in Zoom that fits both you and me.

Attendance and Participation

Due to the online nature of the course and the wide range of topics that we will cover, frequent engagement will be essential for your success. Although it is *not* officially a part of the course grade, participation *is* important: long gaps in participation could adversely affect your grade by impacting your understanding of the material. It can be tempting to view video lectures as a substitute for taking notes but a significant body of research suggests that taking notes by hand improves learning outcomes for students.

That said, I understand that you may occasionally have difficulties keeping up with the pace of the course. Please communicate with me in advance if possible, so I can help point you to useful resources.

Expectations for Student Effort

You are expected to spend a minimum of 9 hours per week for an online 3-credit course engaged in activities including, but not limited to: reading, listening to/viewing media, studying for and completing assignments and reviewing instructor feedback, contributing to discussions, conducting research etc.

Students enrolled in online courses are subject to the same academic regulations as on-campus students. For the most accurate and up to date information go to Academic Regulations.

Student Learning Outcomes

Students who successfully complete the course will be able to:

- describe different types, uses, and structure of data sets;
- perform basic procedures to obtain, process (clean), and store data;
- understand and compute simple summary statistics and statistical models;
- construct simple scripts for processing, analyzing, and visualizing data;
- conduct exploratory data analysis;
- apply elementary (supervised and unsupervised) learning techniques;
- analyze published data analytics work across multiple application domains;
- increase the level of the professional preparation, namely:
  - work collaboratively on data analytics projects;
  - present data preparation processes and the results of analyses;
  - understand legal and ethical ramifications of data-driven projects;
  - curate and store data sets;
  - use common programming tools and computational platforms.

Assignments and Assessments

The total number of points you can get in this course is 100 (plus up to 2 extra credit points), so you can treat points from the course assignments and assessments as percentages of your overall course grade.

- **Reading and Discussion Assignments (RDA):** These assignments involve completing several short readings on a relevant data analytics topic. A set of discussion questions related to the reading will be provided and you will be required to make a discussion post responding to those questions. These are more subjectively graded and will receive a score - the main goal here is participation. There will be 13 RDAs in total, each worth 1 point, and only 10 highest scores will count towards the course grade. Late submissions are allowed but they are subject to a possible late penalty at my discretion.

- **Problem Sets:** These assignments will usually be a mixture of direct questions about the lecture material and opportunities for you to apply the methods we discuss to real data. There will be 6 problem sets in total, each worth 5 points, and the highest score will count twice towards the course grade. Submit your work on Canvas before the due date specified, late submissions will not receive credit (except for the prior accommodations). You are required to submit your work both as a completed Jupyter notebook file and as a .html file. You will work on problem sets in the teams of 3-5 students. Only one submission from a team is expected; however, include only the names of the students who *actually contributed* to the team work on the current problem set. Students who failed to put work on a problem set will receive zero credit for that assignment.
  Late submissions will not receive credit. After receiving your graded work back, I allow you to resubmit your corrected work to receive 50% of the deducted points back. This applies to a missing homework as well, i.e. if you missed the original deadline of a problem set (or did not contribute to your team's work thus earning zero points), then you can resubmit it to receive a partial credit (the highest possible value of the resubmitted work will be half of the original points of the assignment). Resubmission of the corrected work is done *individually* by attaching files (completed Jupyter notebook file and .html file) as a comment on the Canvas page for the problem set. Note that each student who is willing to improve their score for a problem set needs to resubmit their updated work independently from their teammates.

- **Quizzes:** These individual assignments cover theoretical concepts and generally consist of 10 multiple choice questions. There will be 5 quizzes in total, each worth 2.5 points, and the highest score will count twice towards the course grade. Quizzes are timed to be completed in 60 minutes or less.

- **Midterm Exam**: There will be a single exam at the end of Week 7, worth 15 points. This individual assignment is an open notes exam, and it will be submitted on Canvas. The exam is timed to be completed in 120 minutes or less.

- **Final Project:** The final assessment in the course will be a project, where you will complete a data analytics task from beginning to end. More details will be discussed later in the semester, but this will provide you with flexibility to tackle a topic of broader depth than those encountered in the problem sets and quizzes. You will work in the teams of 3-5 students on your final project (same teams as for the problem sets). The final project is worth 25 points. There are three deliverables regarding the project: a completed team contract (worth 1 point) is due the end of Week 2, a project proposal (worth 2 points) is due the end of Week 8, and a project report/presentation (worth 12 points) due the end of the Finals Week. More details will be provided later in the course.

  Remaining 10 points will be assigned based on a peer evaluation, where you will anonymously evaluate the contribution of your teammates to problem sets and the final project; 2 points will be given for filling in an evaluation form, and up to 8 points will be given based on the evaluation feedback from your teammates. Peer evaluation is due the end of the Finals Week.

- *Extra credit (optional):* There will be two optional surveys: an introduction survey is due the end of Week 1 and an exit survey is due the end of the Finals Week. Both surveys are worth 1 point each.

Grading Policy

The "weights" of course assignments towards the overall course grade is as follows:

Assignment	Percentage
RDA	10%
Problem Sets	35%
Quizzes	15%
Midterm Exam	15%
Final Project	25%
Extra credit (optional)	*2%*

I will determine your letter grade using the following grade schema:

A	93-100
A-	90-92.99
B+	87-89.99
B	83-86.99
B-	80-82.99
C+	77-79.99
C	73-76.99
C-	70-72.99
D+	65-69.99
D	60-64.99
F	0-59.99

Late Work Policy

Late submissions are not normally allowed (except for RDAs, where they are subject to a late penalty). Earlier submissions are allowed at any time before the due date. Extensions *may be* allowed by contacting me well in advance and in case of real emergencies. After the answers for the assignments have been provided on the Canvas website, that problem set or exam cannot be made up, so please reach out to me early if needed. It is better to turn in a partially completed work than none at all.

For an excused missed midterm exam, the accommodation is at my sole discretion and may include a reweighting of the remaining components making up the student's grade or taking a makeup exam. Again, please reach out to me well in advance.

Collaboration Policy

You are encouraged (and sometimes required) to work with other students for the assignments in the class. However, the work that you submit for the *individual* assignments should be your own and in particular should be written in your own words and communicate your own understanding of the solution. If you do collaborate, please list the names of the other students you worked with on your submission. You may be asked to explain your work in person to obtain full credit. Obtaining solutions for course problems from external sources will be considered a violation of the academic integrity policy with consequences described below.

Online Discussion Policy

The essence of education is exposure to diverse viewpoints. In your discussion posts you'll meet students with vastly different opinions and backgrounds. You're encouraged to disagree with the substance of others' ideas and opinions but do so with an active sense of respect for one another, and without losing focus on the topic at hand. Personal attacks, inflammatory statements, flaming, trolling, and disruption of the discussion do not have a place in academic discourse. Postings must comply with University policy on use of computing resources, including those regarding harassment and discrimination, as well as conform to the WSU Community Standards.

I will aim to promote high-quality academic discussions by removing any posts I view as disruptive of the educational process and alerting students whose posts have been removed that they have violated course expectations. Students who continue to misuse the discussion boards after a warning may be subject to removal of access rights, course failure, and referral to the Office of Community Standards. Visit WSU Netiquette Guidelines.

Academic Integrity

Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course.

You are responsible for reading WSU's Academic Integrity Policy, which is based on Washington State law. Students who violate the Academic Integrity Policy will fail the assignment, will not have the option to withdraw from the course pending an appeal, and will be reported to the Center for Community Standards. Multiple violations of the policy will cause you to fail the course.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students (identified in Washington Administrative Code (WAC) 504-26-010(2)). You need to read and understand all of the definitions of cheating. If you have any questions about what is and is not allowed in this course, please reach out to me before proceeding.

If you wish to appeal my decision relating to academic integrity, please use the form at the Center for Community Standards website. You must submit this request within 21 calendar days of the decision.

Incomplete Grade Policy (Academic Rule 90h)

Incompletes are granted only with my permission and are subject to the following guidelines:

1. You must request an incomplete in writing or by e-mail to me before the end of the semester. This request must be signed and dated by you (or identified by your WSU e-mail address) and must explain the reasons behind the request for the incomplete.
2. In order to be considered for an incomplete grade, these two conditions should be met:
    - You must complete a minimum of 75 percent of the assigned course work.
    - You must have a mathematical possibility of scoring a 60 percent or above for the entire course.
3. If extraordinary circumstances (e.g., family emergency, serious illness) are involved, I retain the discretion to grant an incomplete even if the minimum conditions outlined in item 2 above are not met.

If an incomplete grade is granted, the standard WSU policy applies (i.e., ALL work must be completed within one full year from the end of the enrollment semester at issue, unless a shorter time is specified by the instructor. Otherwise, an automatic grade of "F," or failing, will be entered on your transcript).

Library Support

All students enrolled in Washington State University online courses can use the WSU Libraries online databases and receive reference and research assistance from their home campus. Students can also borrow books and other circulating material as well as access full-text journal articles.

General Library Support Links:

- Global Campus
- Pullman Campus
- Spokane Campus
- Tri-Cities Campus
- Vancouver Campus
- College of Nursing

Online Tutoring

As a WSU student enrolled in an undergraduate course, you have free unlimited access to Online Tutoring. This is not a course requirement, but a resource for you to utilize as needed.

With three ways to access a tutor you can choose the one that best fits your needs:

- Submit a paper: Writing Lab tutors will respond to papers in any academic subject. Just submit your paper, ask specific questions on the submission form, and a tutor will respond within 24-48 hours.
- Live tutoring: eChat rooms allow students to meet with tutors in one-on-one tutoring sessions via a fully interactive, virtual online environment.
- Leave a question: Students can leave specific questions for a tutor in any of our subjects by taking advantage of our eQuestions option. Our tutors will respond to your question within 24-48 hours.

More details and the list of available tutoring subjects can be found at www.eTutoringOnline.org

WSU Academic Calendar

Please refer to the WSU academic calendar to be aware of university holidays and important deadlines throughout the semester.

Technical Support

The WSU Global Campus Current Student website has all the administrative and contact-related information you need to be a successful online learner as well as study tips and skills for success.

Before contacting Technical Support please visit our Equipment Requirements to ensure your system meets our computer requirements and test your connection.

Technical support is available 24/7. A member of our technical support team will assist you or direct your request to the appropriate group. The technical support team is NOT able to answer questions about course content or procedures (e.g., assignments, grades).

Please visit our Tech Support website for hours of operation and contact information.

Copyright

University Syllabus

Students are responsible for reading and understanding all university-wide policies and resources pertaining to all courses (for instance: accommodations, care resources, policies on discrimination or harassment), which can be found in the University Syllabus