HCAI5DS02 - Data Analytics and Visualization. Unlocking Decisions through Data: Where Analytics Meets Storytelling and Statistical Insight.

Prepared By: Siman Giri

Summer, 2025

1 Basic Course Information.

- **Promoter: Siman Giri {siman.giri@herald}.**
- **Instructors:**
  - **Ronit Shrestha.**
  - **Shiv Kumar Yadav.**
- **Lectures: TBA.**
- **Office Hours: On Demand and By Appointment.**
- **Quick Links and Navigation:**
  - **Course Logistics and Learning Outcomes.**
  - **Final Grading.**
  - **Getting Help.**
  - **Academic Integrity Policies.**
  - **Approximate Schedule.**

2 Course Logistics.

2.1 Course Introduction:

This course is a component of the **Practical Data Science curriculum**, developed and offered by **Herald College – Center for AI** as part of its ongoing Education Initiative. It is specifically designed for **Level 5 students** who have recently completed or are awaiting their results.
The course covers key techniques used in real world data science applications, including:

- **Statistical Analysis:** Formulate hypotheses and perform statistical analysis on real-world datasets to derive actionable insights.
- **Regression & Causation:** Understand relationships in data using correlation and regression techniques, and explore causal inference in applied settings.
- **Robust Linear Models:** Build and evaluate linear models that are resilient to outliers and violations of standard assumptions.
- **Time Series Forecasting:** Model and forecast time series data using moving average and autoregressive models, with applications in domains like finance.
- **Communication:** Clearly communicate analytical results through visualizations, structured reports, and professional presentations.

A central objective of this course is to cultivate **a critical understanding of data-driven analysis, visualization design, and interpretability**. Students will explore the influence of**data preprocessing, exploratory data analysis, and statistical modeling** on the quality of insights drawn from data. The course also emphasizes the significance of **transparency, reproducibility, and ethical data communication** in modern analytical practices.

2.2 Pre-requisites:

Must have obtained passing grades on following Modules:

- Introduction to Python Programming - Level - 4.
- Computational Mathematics - Level - 4.
- 5CS037 - Concepts and Technologies of AI - Level - 5.

2.3 Recommended Reading:

There are no required textbooks for this course. Lecture slides, notes, and worksheets will be sufficient for covering the course material. If additional reading materials or research papers are needed, they will be provided as necessary. However, the following resources are recommended to further support your learning:

1. **Naked Statistics: Stripping the Dead from the Data**
   Charles Wheelan
   A friendly introduction to statistics using real-life examples and storytelling.

2. **Storytelling with Data: A Data Visualization Guide for Business Professionals**
   Cole Nussbaumer Knaflic
   Teaches how to communicate data clearly through effective visual storytelling.

3. **Fundamentals of Data Visualization**
   Claus O. Wilke
   Covers essential design principles for creating clear, informative, and visually appealing data graphics.
   [**Free Online**] `https://clauswilke.com/dataviz/`

4. **Think Stats: Probability and Statistics for Programmers**
   Allen B. Downey
   Introduces statistical concepts using Python and practical examples.
   [**Free PDF**] `https://greenteapress.com/wp/think-stats/`

5. **An Introduction to Statistical Learning (ISL)**
   Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
   Provides practical guidance on statistical learning methods with R examples.
   [**Free PDF**] `https://www.statlearning.com/`

2.4 Learning Outcomes:

Upon successful completion of this course, students will be able to:

1. Apply Foundational Statistical Techniques:
    - Demonstrate a solid understanding of core concepts in descriptive and inferential statistics.
    - Formulate hypotheses and perform statistical analysis on real-world datasets.
2. Develop Competence in Data Wrangling and Visualization:
    - Clean, transform, and prepare datasets for analysis using Python libraries such as Pandas and NumPy.
    - Create meaningful visualizations using tools like Matplotlib, Seaborn, and Plotly to support data-driven storytelling.
3. Perform Advanced Analytical and Modeling Tasks:
    - Apply regression analysis and explore causal relationships in structured data.
    - Analyze temporal patterns and forecast time series using moving averages and autoregressive models.
4. Communicate Insights with Clarity and Rigor:
    - Present findings through interactive dashboards, written reports, and visual presentations.
    - Critically evaluate the ethical implications of data analysis and visualization choices.

3 Final Grading:

The requirements for this course include active participation in class discussions, completion of assigned lab exercises, a mid-term quiz, and the final course project. The grading breakdown is as follows:

- 10% **Quiz:** A quiz in MCQ format, Date and Time to be announced.
- 25% **Weekly Worksheets, Problem Sets and Question Answer.**
- 60% **Final Course Project:** The final project should comprehensively demonstrate the knowledge and skills acquired throughout the course. It is expected to integrate key concepts from statistical analysis, data visualization, and practical data science techniques to solve a real-world problem or present insightful findings from a dataset.
- 5% **Participation:** Active participation will be assessed through in-class quizzes and involvement in class discussions. Please note that no make-up quizzes will be provided for missed classes.

This grading structure is designed to encourage consistent engagement with the material and foster both individual and collaborative learning throughout the course.
Caution: These plans are tentative and subject to change. Final confirmation will be provided.

4 Getting Help:

The preferred method of communication for course-related inquiries is via email, ensuring a timely response. For urgent matters or recommendations, students are encouraged to meet with the teaching staff or the designated teaching assistant (TA) during their office hours.

4.1 About Office Hours:

The teaching staff holds weekly office hours to assist students with course-related matters. Students can obtain details of office hours time by contacting their **designated instructor**, the **Student Support Desk (SSD)**, or the **Personal Academic Tutor (PAT) office**. Meetings outside of scheduled office hours are strictly by prior **appointment only**.

5 Academic Integrity Policies:

The Academic Integrity Policy of this course must be strictly adhered to when completing assignments and participating in discussions.**Please - Read this carefully.**

5.1 Collaboration among Students:

Collaboration among students is intended to **facilitate deeper learning and comprehension, rather than to circumvent the learning process**. Engaging in group discussions and collaborative study of course materials is strongly encouraged. Students may seek clarification and conceptual guidance from their peers to enhance their understanding of the subject matter required for completing assignments. However, such collaboration must support independent learning rather than substitute for individual effort. **The direct reproduction of any material from other students or external sources is strictly prohibited**. All submitted work must be the sole effort of each student.
All of the following activities will be considered cheating:

1. **Sharing or Copying code, files or answers:** whether through direct copying, retyping or using online sources e.g. Stack-overflow Git-hub without proper attribution.
2. **Submitting work that is not original:** including using external code or code generated by LLMs, with intention of passing off such work as the student's own.
3. **Copying answers** to quizzes, assignments, or projects from another individual or from any published or unpublished written or electronic sources.
4. **Collaborating with others** on individual assignments, quizzes or project without explicit permission from the instructor.

5.2 Duty to Protect One's Work:

Students are responsible for safeguarding their work against unauthorized access, copying, or misuse by others.If a student's work is copied by another student, both parties will be held accountable for violating course policies. This applies regardless of whether the original author explicitly permitted the copying or failed to take adequate precautions to prevent it. If identical or highly similar work is submitted by multiple students, all involved will face academic penalties.

To uphold academic integrity and protect future students, solutions to assignments must not be shared publicly, either during the course or after its completion.

5.3 Duty to Give Viva:

It is **student's responsibility to attend viva** examinations at the scheduled time determined by the instructor. Failure to comply with this requirement may result in academic penalties.

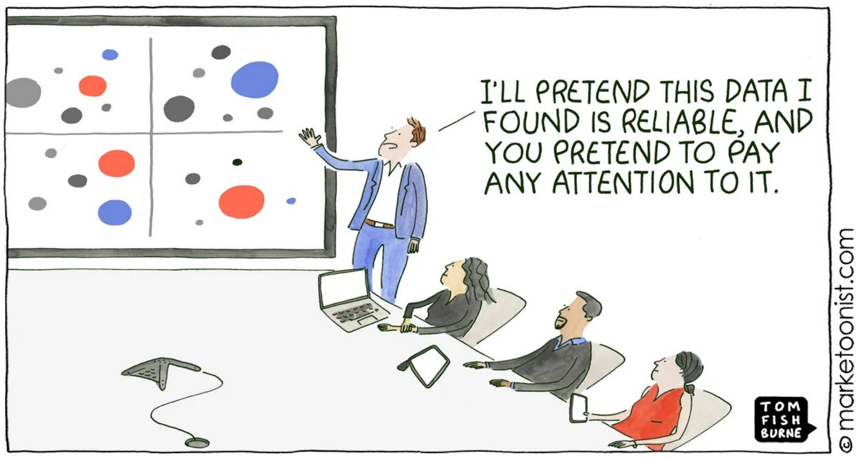
5.4 Plagiarism and AI - Generated Content:

Plagiarism, including the use of AI-generated content without proper attribution, constitutes a violation of academic integrity. Students must ensure that all work submitted for evaluation is their own and properly cites any external sources, including AI-generated material, used in their assignments.
Using AI tools to generate content without disclosing their use, or presenting AI-generated work as one's own, is considered plagiarism. Any attempt to submit plagiarized work, whether through manual copying or the use of AI tools, will be subject to academic penalties. It is essential that students recognize the importance of maintaining transparency regarding the sources and tools utilized in their work to uphold academic standards.

5.5 Other General Policy:

Students are expected to **attend all lectures, tutorials, and workshops** regularly. **Active participation** in class discussions is encouraged, and students must ensure the timely submission of all coursework as assigned by the instructor. While late submissions may be accepted under certain circumstances, they will incur penalties as per the course policy.

In addition to the aforementioned policies, students are expected to adhere to all rules and regulations established by the College . **Failure to comply with these policies may result in severe academic penalties, including potential exclusion from the module.**


I'LL PRETEND THIS DATA I FOUND IS RELIABLE, AND YOU PRETEND TO PAY ANY ATTENTION TO IT.
TOM FISH BURNE
© marketoonist.com

6 Approximate Schedule

Here you will find a detailed breakdown of the weekly topics and instructional themes for the course. Please note that this proposed schedule is tentative and may be adjusted to accommodate pedagogical priorities and the pacing of instruction.

Week 1: Foundations of Statistical Thinking:

Theme: Understanding Data through Descriptive Statistics and Visualization

Topics Covered

- The role of statistical thinking in data analytics and decision-making
- Understanding the data generating process: observational vs. experimental data
- Descriptive statistics - Numerical and Graphical Summaries:
- How to summarize and interpret real-world data using Python
- Case studies and discussion: interpreting data stories using descriptive statistics

Tools / Libraries:

❑ `Python` ❑ `Pandas` ❑ `Matplotlib` ❑ `Seaborn` ❑ `Jupyter Notebooks`

Learning Objectives:

By the end of this week, students will be able to:

- Explain the importance of statistical thinking in understanding data ✓
- Differentiate between types of data and recognize common biases in data collection ✓
- Calculate and interpret key descriptive statistics using Python ✓
- Construct and critique visual summaries of real-world datasets ✓
- Apply Python libraries to analyze and visualize structured datasets ✓
- Interpret data narratives through exploratory analysis and case studies ✓

Lab sheet - Week - 1 - Understanding Data with Descriptive Statistics.

Week 2: Fundamentals of Probability Theory:

Theme: Understanding and Quantifying Uncertainty

Topics Covered:

- What is probability? Classical, empirical, and subjective interpretations
- Sample space and events
- Rules of probability: addition and multiplication rules
- Conditional probability and independence
- Bayes' Theorem: intuition and applications
- Real-life examples: weather forecasting, fraud detection, recommendation systems

Activities:

- Simple probability simulations using Python
- Calculating conditional probabilities from data
- Short Exercise: interpreting Bayes Theorem

Tools / Libraries:

❑ `NumPy - Random Module` ❑ `Matplotlib` ❑ `Seaborn` ❑ `Jupyter Notebooks`

Learning Objectives:

- Explain different interpretations of probability and apply them in context ✓
- Identify and compute probabilities for events within a sample space ✓
- Apply conditional probability and the concept of independence to real-world data ✓
- Use Bayes' Theorem to update beliefs based on new evidence ✓
- Simulate probabilistic events and visualize outcomes using Python ✓
- Analyze and interpret probabilistic scenarios using tools like Venn and tree diagrams ✓

Lab sheet - Week - 2 - Simulating and Interpreting Probability in Python.

Week 3: Random Variables and Probability Distributions

Theme: Modeling Real-World Uncertainty with Distributions

Topics Covered

- Discrete and continuous random variables
- Probability mass functions (PMFs) and probability density functions (PDFs)
- Cumulative distribution functions (CDFs)
- Common distributions:
  - Discrete: Bernoulli, Binomial, Poisson
  - Continuous: Uniform, Normal, Exponential
- Expected value, variance, and standard deviation

Tools / Libraries:

❑ `SciPy (stats)`

Learning Objectives

- Differentiate between discrete and continuous random variables ✓
- Describe and compute PMFs, PDFs, and CDFs for key probability distributions ✓
- Interpret properties of distributions such as expected value and variance ✓
- Recognize and model real-world phenomena using Bernoulli, Binomial, Poisson, Normal, and Exponential distributions ✓
- Visualize and compare distributions using Python libraries ✓
- Apply appropriate distributions to answer probabilistic questions using real datasets ✓

Lab sheet - Week - 3 - Exploring Discrete and Continuous Distributions in Python.

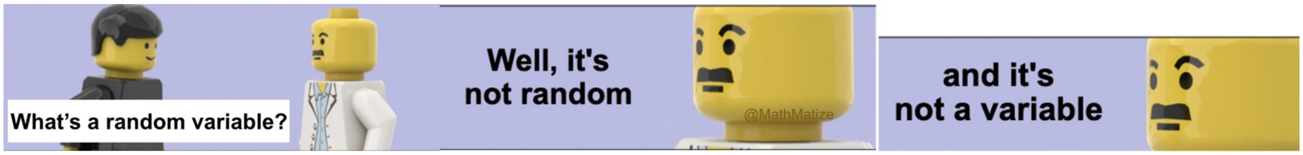


Figure 1: Does There Exist a Randomness?

Week 4: Statistical Inference and Hypothesis Testing:

Theme: Drawing Conclusions from Data — Confidence Intervals and Hypothesis Tests

Topics Covered:

- The logic of statistical inference from sample to population
- Constructing and interpreting confidence intervals
- Hypothesis testing: t-tests, chi-square, and proportion tests
- The scientific method in data analysis
- Real-world data analysis using Python's `scipy.stats`

Tools / Libraries:

❑ `SciPy (stats)`

Learning Objectives:

- Understand the logic of statistical inference ✓
- Construct and interpret confidence intervals ✓
- Perform hypothesis tests using appropriate statistical tests ✓
- Apply the scientific method in analyzing data ✓
- Analyze datasets using `scipy.stats` ✓

Lab sheet - Week - 4 - Performing Inference on Real World Dataset.

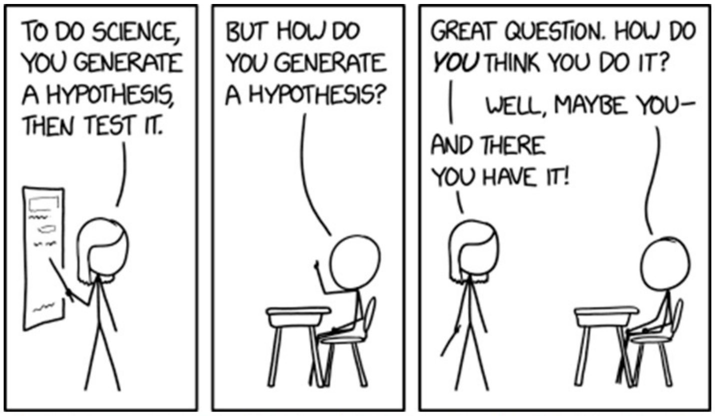

TO DO SCIENCE, YOU GENERATE A HYPOTHESIS, THEN TEST IT.
BUT HOW DO YOU GENERATE A HYPOTHESIS?
GREAT QUESTION. HOW DO *YOU* THINK YOU DO IT?
WELL, MAYBE YOU–
AND THERE YOU HAVE IT!

Image From: xkcd comic 2569 (Hypothesis generation) by Randall Munroe.

Figure 2: Testing Your Hypothesis.

Week 5: Correlation, Regression, and Causality:

Theme: Quantifying Relationships and Exploring Causal Effects

Topics Covered

- Covariance and correlation: computation and interpretation
- Simple linear regression and coefficient interpretation
- Regression assumptions and residual analysis
- Limits of correlation and regression in inferring causality
- Observational vs. experimental data

Tools / Libraries:

❑ `statsmodels`

Learning Objectives:

- Compute and interpret covariance and correlation ✓
- Build and interpret simple linear regression models ✓
- Check regression assumptions using residual plots ✓
- Understand when and how causality can be inferred ✓

Lab sheet - Week - 5 - Correlation vs Causation Analysis and Diagnostic Plot.

EVERYBODY WHO WENT TO THE MOON HAS EATEN CHICKEN!
GOOD GRIEF. CHICKEN MAKES YOU GO TO THE MOON!

Image From Internet: Subject to Copyright.

Figure 3: Eat your Chicken !!!

Week 6: Designing Experiments and A/B Testing:

Theme: Causal Inference through Controlled Experiments

Topics Covered

- Principles of experimental design
- Setting up and simulating A/B tests
- Statistical vs. practical significance
- Recognizing biases and confounding variables
- Analyzing and interpreting test results

Tools / Libraries:

❑ `scipy.stats`

Learning Objectives:

- Understand and apply experimental design principles ✓
- Conduct and interpret A/B testing simulations ✓
- Distinguish between statistical and practical significance ✓
- Identify potential biases and confounding variables ✓
- Analyze real experimental data for valid inferences ✓

Lab sheet - Week - 6 - Designing and Simulating A/B Tests.

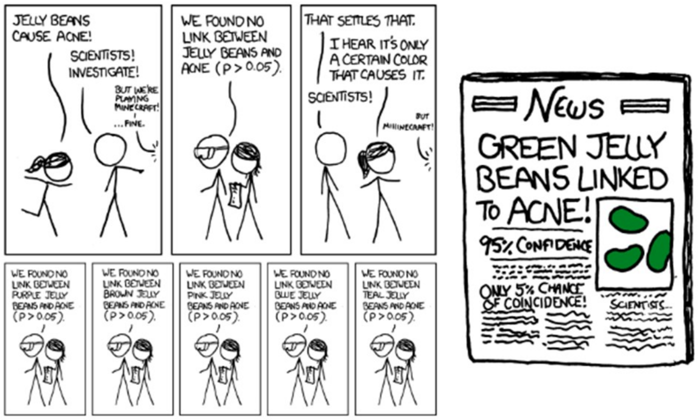

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Image From: xkcd comic 2569 (Hypothesis generation) by Randall Munroe.

Figure 4: Take Care of Your Skin !!!

Week 7: Introduction to Time Series Analysis and Forecasting

Theme: Temporal Data Analytics

Topics Covered:

- Time series structure: trend, seasonality, noise
- Moving averages and smoothing techniques
- Autocorrelation and partial autocorrelation
- AR, MA, and ARIMA models (introductory)
- Applications: finance, economics, environment

Tools / Libraries:

❑ `statsmodels` ❑ `sklearn`

Learning Objectives:

- Understand components of time series ✓
- Use moving averages to smooth data ✓
- Analyze autocorrelation patterns ✓
- Fit basic AR/MA models ✓
- Apply time series techniques on real datasets ✓

Lab sheet - Week - 7 - Visualizing and Forecasting Time Series in Python.

AND AFTER BIG DATA SOLVES ALL OUR PROBLEMS, WE'LL RIDE AWAY ON MAGIC FLYING UNICORNS.
FORECAST
© marketoonist.com

Figure 5: Profitsssss !!!

Week 8: Visualizing Data — Principles and Practice:

Theme: Effective Communication Through Visual Representation

Topics Covered:

- Perceptual principles and cognitive biases
- Choosing chart types: bar, line, scatter, pie, etc.
- Design elements: color, layout, labels, annotations
- Critiquing and redesigning poor visualizations
- Telling data stories

Tools / Libraries:

❏ `matplotlib` ❏ `seaborn` ❏ `pandas`

Learning Objectives:

- Apply design principles for clear visualizations ✓
- Choose appropriate chart types based on data ✓
- Avoid misleading visuals through best practices ✓
- Use Python libraries for static data visualization ✓

Lab sheet - Week - 8 - Redesigning Visuals and Applying Visualization Best Practices.

Figure 6: Dataaaa !!!

Week 9: Interactive Visualization with Plotly and Dash

Theme: Bringing Data to Life with Interactivity

Topics Covered

- Introduction to Plotly and Dash
- Creating interactive charts (line, bar, scatter)
- Dashboard layout and filtering
- Dynamic user input controls
- Hosting and sharing simple apps

Tools / Libraries:

❑ `Plotly` ❑ `Dash` ❑ `pandas` ❑ `Jupyter Notebooks`

Learning Objectives:

- Build interactive visualizations using Plotly ✓
- Create dashboards with Dash ✓
- Filter and highlight data dynamically ✓
- Share and deploy mini analytics apps ✓

Lab sheet - Week - 9 - Building Interactive Dashboards with Plotly and Dash.

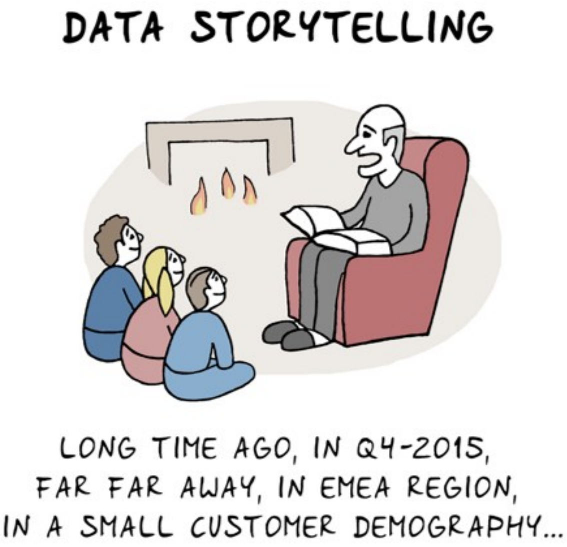

DATA STORYTELLING
LONG TIME AGO, IN Q4-2015,
FAR FAR AWAY, IN EMEA REGION,
IN A SMALL CUSTOMER DEMOGRAPHY...

Figure 7: Moreeeee - Dataaaa !!!

Week 10: Real-World Visualization Examples — Geospatial and Sports Data

Theme: Domain-Based Applied Visualization

Topics Covered

- Working with geospatial data and maps
- Visualizing location-based data with Folium and Plotly
- Choropleth, scatter maps, density maps
- Sports analytics and performance dashboards
- Radar charts, rolling averages, comparison visualizations

Tools / Libraries:

❑ `GeoPandas` ❑ `Folium` ❑ `Plotly`

Learning Objectives

- Work with geospatial data in Python ✓
- Visualize and analyze location-based data ✓
- Build sports performance dashboards ✓
- Apply advanced chart types and comparisons ✓

Lab sheet - Week - 10 - Mapping and Analyzing Sports and Geospatial Data.

18
19
2
7
3
Geft
"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

Figure 8: Goooooal !!!

The - End