The Hong Kong University of Science and Technology (Guangzhou)

UG Course Syllabus

[Course Title] Data Analysis with Python

[Course Code] FTEC 3210

[No. of Credits] 3

[Any pre-/co-requisites] DSAA 2011 Machine Learning

Name: Sijia Chen

Email: sijiachen@hkust-gz.edu.cn

Office Hours: Room W4-521 / Anytime after email contact

Lecture (Tutorial) Day, Time, and Venue

Tue/Thu 13:30–16:30 (Summer 2026, June 16 – July 28);

Venue: W3-105.

Course Description

FTEC 3210 is an intensive summer course built on a radical premise: every stage of the data-analysis, coding, and academic-reproduction workflow is completed through a fully AI-driven paradigm, with Large Language Models (LLMs) acting as the primary co-pilot. Moving beyond syntax memorisation, the course focuses on the complete lifecycle of computational research — from parsing a frontier paper, generating and debugging code, running rigorous statistical analyses, to writing an academic manuscript. Students learn to construct end-to-end pipelines with AI-assisted coding tools (Cursor, Claude Code, Codex, Qoder / QoderWork, Qwen Code, and others), to understand the statistical principles behind each analytical method they invoke, to build and audit predictive models (regression, time series, classification, clustering, self-supervised representation learning), to design multi-agent workflows that autonomously carry out data-analysis tasks, and to leverage frontier auto-research systems to move from an initial idea to a publication-quality manuscript. Ethical use of AI, critical validation of AI-generated outputs, and reproducibility discipline are treated as first-class skills alongside the technical content.

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

1. Adopt an AI-first research paradigm — articulate how a fully AI-driven data-analysis workflow differs from traditional manual programming, and place any modern AI coding tool inside a principled taxonomy (CLI agent, cloud agent, BYOK IDE, workflow framework) based on its Agent Harness H = (E, T, C, S, L, V).

2. Construct end-to-end data pipelines with AI collaboration — drive CLI and workflow-based agents (Cursor, Claude Code, Codex, Qoder / QoderWork, Qwen Code, LangGraph, CrewAI, Coze Studio) through fetch →

clean → feature-engineer → visualise → report on real financial datasets, producing a reproducible repository backed by SKILL.md and AGENTS.md specifications.

3. Apply and critically validate statistical methods on real data — build a full statistical profile (descriptive statistics, hypothesis testing, correlation analysis), understand the assumptions of each test, and audit AI-generated statistical code for wrong-test errors, p-hacking, look-ahead bias, and correlation-versus-causation confusion.

4. Build, interpret, and audit predictive models — engineer features and fit regression, ARIMA, and modern classification models (Logistic Regression, tree ensembles, XGBoost / LightGBM / CatBoost, TabPFN) with walk-forward cross-validation; interpret ROC / PR-AUC / calibration / SHAP outputs; and recognise when a signal is genuinely un-predictable.

5. Design autonomous data agents and multi-agent workflows — assemble agent frameworks (LangGraph, MCP, OpenAI Agents SDK, Anthropic patterns) that plan, execute, and self-correct multi-step data-analysis tasks, and evaluate when a single agent, workflow, or multi-agent orchestration is the appropriate design choice.

6. Deliver a research artefact through an AI-driven auto-research pipeline — leverage frontier systems (AI Scientist, Agent Laboratory, AI co-scientist, autoresearch, DeepAnalyze) alongside the AgenticFib Lab pipeline to perform literature review, experimental design, execution, statistical validation, publication-quality figure generation, and manuscript drafting for a reproducible baseline plus a novel modification — while disclosing AI collaboration transparently.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics for each assignment are provided below, outlining the criteria used for evaluation.

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage, and due dates; perhaps, add a summary table as below, to precede the details for each assessment.]

Assessment Task	Contribution to Overall Course grade (%)	Due date
Assignment 1: Paper Deconstruction & Reproduction Blueprint (B+ target)	27%	Tue, Jun 30, 2026 (23:59)
Assignment 2: AI-Driven Reproduction & Baseline Validation (A– / A target)	29%	Thu, Jul 16, 2026 (23:59)
Final Assignment: Paper Improvement & Performance Enhancement Analysis (A+ target)	29%	Tue, Jul 28, 2026 (23:59)
Final Assignment: Presentation	15%	Tue, Jul 28, 2026 (23:59), but the deadline will be extended as needed until all students have participated.

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

[add to/delete table as appropriate]

Assessed Task	Mapped ILOs	Explanation
Assignment 1	ILO 1, ILO 2, ILO 6	The blueprint assesses students' ability to adopt the AI-first paradigm (ILO 1), use AI coding tools to construct a reproducible pipeline plan from a frontier paper (ILO 2), and disclose their AI collaboration transparently in prompt / iteration logs (ILO 6).
Assignment 2	ILO 2, ILO 3, ILO 4	The reproduction task assesses students' ability to drive AI agents through an end-to-end pipeline (ILO 2), apply and audit statistical methods rigorously (ILO 3), and build/interpret the paper's baseline predictive models with walk-forward validation (ILO 4). Every headline number must be traceable to a re-runnable script.
Final Assignment (Project + Presentation)	ILO 3, ILO 4, ILO 5, ILO 6	The final assignment assesses students' ability to critically validate AI-generated statistics (ILO 3), design and diagnose a principled model improvement with ablation (ILO 4), orchestrate autonomous agents inside the improvement loop (ILO 5), and deliver a manuscript-style report through an AI-driven auto-research pipeline (ILO 6).

Grading Rubrics

[Detailed rubrics for each assignment will be provided. These rubrics clearly outline the criteria used for evaluation. Students can refer to these rubrics to understand how their work will be assessed.]

Final Grade Descriptors:

[As appropriate to the course and aligned with university standards]

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Consistently exceeds expectations across all ILOs. Delivers a fully reproducible AI-driven pipeline, a rigorously validated baseline, and a principled novel modification with credible performance gains and honest ablation analysis. Prompt / iteration logs read as a case study of expert AI orchestration; statistical audit catches

		every major failure mode (data leakage, wrong CV, p-hacking, fabricated numbers). Written communication is publication-adjacent.
B	Good Performance	Meets every ILO competently. Baseline reproduction is correct and the improvement is real but modest; the audit of AI outputs catches most but not all failure modes. Minor gaps in depth, novelty, or reproducibility, but the work is trustworthy end-to-end and the AI-collaboration log is disciplined.
C	Satisfactory Performance	Adequately addresses core ILOs. Standard analyses and visualisations; the pipeline runs and the baseline is close but not fully aligned with the paper. Some AI outputs are accepted uncritically; multi-agent workflow is functional but not autonomous. The improvement is superficial or its evaluation is incomplete.
D	Marginal Pass	Partially fulfils ILOs. Data integration or statistical validation is incomplete; models are built but not diagnosed; AI-generated code is used without meaningful audit. AI-collaboration log is thin or missing. Threshold-level engagement.
F	Fail	Fails to meet essential ILOs. Baseline reproduction is unsuccessful; statistical claims are unsupported or fabricated; there is no evidence of critical AI validation or of independent methodological understanding. Suspected undisclosed AI usage is treated as academic misconduct.

Course AI Policy

This course **actively encourages and welcomes** the use of AI tools throughout every stage of the work — AI-powered coding assistants, LLM-based literature review agents, automated statistical-analysis workflows, and multi-agent frameworks. Rather than treating AI as a convenience, this course integrates it as a core competency: the entire course is designed on the premise that a fully AI-driven paradigm is now the professional standard. Students are expected to use AI tools thoughtfully, creatively, and transparently — to enhance, not replace, their own analytical reasoning and technical judgement. Every submission must (i) disclose AI usage in full, attaching the prompt / iteration log and naming the tools used; (ii) critically validate AI outputs, never accepting AI-generated code, statistical claims, citations, or numbers without independent verification; (iii) maintain scholarly integrity, treating AI as a disclosed collaborator; and (iv) include a reflection identifying at least three concrete moments where the AI failed and how the student intervened. Undisclosed AI usage and AI-hallucinated citations or results are treated as academic misconduct.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within one week of submission. Students who have further questions about the feedback including marks should consult the instructor within 3 working days after the feedback is received.

Resubmission Policy

Each assignment may be assessed only once, underscoring the importance of diligence and thoroughness in the initial submission. Students who seek clarification or have concerns about their exam scores must contact the instructor within three working days of receiving their grades. Timely communication enables prompt review and resolution of any discrepancies or misunderstandings.

The decision to offer a reassessment opportunity rests solely with the instructor. This discretion allows consideration of individual circumstances, including the nature of potential errors in the original assessment, the student's overall performance and engagement in the course, and institutional policies. It reflects the instructor's dual role: ensuring fair and consistent evaluation while fostering a supportive learning environment in which students' concerns are heard and addressed appropriately.

Required Texts and Materials

To support your learning throughout the course, I highly recommend the following free, high-quality references:

1. **Python Data Science Handbook**, by Jake VanderPlas – an excellent, hands-on guide covering NumPy, pandas, Matplotlib, and scikit-learn with Jupyter notebooks.
2. **Practical Statistics for Data Scientists**, 2nd Edition.
3. **Kaggle**, https://www.kaggle.com/learn.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST(GZ)'s Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Regulations for Academic Integrity and Student Conduct for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

[Optional] Additional Resources

Free premium LLM access — fully funded accounts and API credits for state-of-the-art LLMs are provided to every enrolled student. Computational reimbursement — cloud-computing or GPU-rental costs incurred by the final project are reimbursable subject to prior approval. Dedicated tech support — direct access to AgenticFib Lab senior researchers and TAs for environment, algorithm, and paper-reproduction issues (supplementary office hours announced on Canvas). Community resources — Kaggle Learn, Papers with Code, Hugging Face, and the Qoder / QoderWork forum for AI coding tool discussions.