The Hang Seng University of Hong Kong

Department of Computer Science

COM6005 Statistical Modelling (2025-2026 Sem 1)

Group Project Guideline

Maximum number of students per group: 5

Important dates (Tentative)**

1. Confirm your project groupmates: **27 Sep 2025 (Sat) 23:59 (Week 4)**
2. PowerPoint file and cover sheet submission deadline: **22 Nov 2025 (Sat) 23:59 (Week 12)**
3. The presentation will be conducted on
   29 Nov 2025 (Sat, lesson time) (Week 13)
   Contingency only: 6 Dec 2025 (Sat, lesson time) (Week 14)
4. Individual report and peer evaluation forms submission deadline: **06 Dec 2025 (Sat) 23:59 (Week 14)**

(I) Description

You are part of an interdisciplinary research team commissioned to develop a data-driven statistical model addressing a topical societal issue relevant to urban communities. Your mission is to apply the complete statistical modeling lifecycle to a real-world problem of your choice that aligns with at least one of the United Nations Sustainable Development Goals (SDGs). These global goals aim to end poverty, protect the planet, and ensure prosperity for all by 2030.

Tasks: Students are required to:

- Identify and define a specific topic with meaningful societal impact that can be addressed using statistical modeling. The topic should relate to at least one SDG, such as Sustainable Cities and Communities (Goal 11), Good Health and Well-being (Goal 3), or Climate Action (Goal 13), etc.
- Source appropriate environmental datasets from open-data platforms such as Kaggle, UCI Machine Learning Repository, data.gov, or Google Dataset Search.
- Design a detailed project title, provide background context emphasizing the SDG relevance, and formulate clear objectives.
- Apply modern statistical modeling techniques learnt in this module—including exploratory data analysis, regression models, classification, and predictive analytics—to analyze your dataset and solve classification or prediction problems.

- Use Python programming in Jupyter notebooks to perform data preprocessing, model building, diagnostic evaluation, and create graphical representations to visualize data patterns and model results.
- Interpret results scientifically and concisely, providing actionable insights to inform policy or intervention strategies that contribute to sustainable societal outcomes.

Example topics might include predicting air quality indices in urban areas to promote SDG 11, modeling healthcare access disparities relating to SDG 3, or analyzing energy consumption patterns for climate action aligned with SDG 13, etc.

Dataset: For this project, you are required to source your own datasets. The following platforms are excellent starting points for discovering high-quality, open-data sources:

1. Open Data on Gov.HK https://data.gov.hk/en/
2. data.gov (from the United States gov): https://data.gov/
3. Kaggle Datasets: https://www.kaggle.com/datasets
4. UCI Machine Learning Repository: https://archive.ics.uci.edu/
5. Google Dataset Search: https://datasetsearch.research.google.com/

(II) Assessment criteria

The total marks for this project are 100. The project assessment criteria include four components:

(1) Group Presentation (30%)

- The presentation **MUST** be conducted in English.
- The length of the presentation is 20 minutes per group (including Question & Answer session)
- Members of each group will be assessed individually.
- Each student must submit the **presentation file** and **cover sheet (Annex 1)** to Moodle before the deadline.
- **After submitting the presentation file, each student needs to submit the topic information to the following Microsoft Form**: https://forms.office.com/r/Tyxr1adxBm

(2) Individual Report (55%)

- The report **MUST** be written in English.
- The report **MUST** include the following sections:
  - Cover page (including the project title, module code and its name, the names and student IDs of the group members, Group number, and name of module instructor)
  - Introduction
  - Methodology and Results
  - Discussion and Conclusion
  - References

- Veriguide report (attach to the end of the report)

- The report MUST follow the following format
  - Page limit: max. 10 pages (excluding Cover page, References, Appendix, and Veriguide report)
  - Page size: A4 size
  - Margin: 1 inch each side
  - Line spacing: Double
  - Font and size: Times New Roman, 12 points
- Each student must submit the **written report** with **Python Program file and Cover sheet (Annex 2)** to Moodle before the deadline.
- Written Report and Python Program files should be compressed into a single ZIP file for submission with a file naming convention: `COM6005_pXXXXXX_Report.zip`, where `pXXXXXX` is your student ID number.

(3) Peer Evaluation (15%)

- Each student must submit a **peer evaluation form (Annex 3)** to the module instructor ***in confidence***.
- In the peer evaluation form, students have to (i) indicate their individual contribution to his/her project, and (ii) give a mark of 0 (lowest) to 10 (highest) **and provide written comments to other group members** to reflect their contribution to the project.
- Each group is required to ask at least one question during the other's presentation. Your group may get at most five marks for the questions.
- **The module instructor will meet with all of the group members individually if any abnormality is observed in the peer evaluation assessment.**
- Each student is required to submit the peer evaluation form (**Annex 3**) to Moodle with the file naming convention: `COM6005_L0Z_GYY_PeerEvaluation_pXXXXXX.pdf`, where `L0Z` is your class session number, `YY` is your group number, `pXXXXXX` is your student ID number.
- Students will receive zero marks if they fail to submit the peer evaluation form on time.

Submission Guide:

Item 1 - Deadline: 22 Nov 2025 (Sat) 23:59 (Week 12) – group and individual submission

- The **group leader** submits only.
- Submit the **PowerPoint file** `COM6005_L0Z_GYY_Presentation.pptx` and the **cover sheet** `COM6005_L0Z_GYY_Report_GroupCoverSheet.zip` (**Annex 1**) to Moodle, where `L0Z` is your class session number, `YY` is your group number in two digits (e.g, Group 6 to be written as `G06`)

- **After submitting the PowerPoint file, <u>each student</u> needs to submit the topic information to the following Microsoft Form**: https://forms.office.com/r/Tyxr1adxBm

Item 2 - Deadline: 06 Dec 2025 (Sat) 23:59 (Week 14) – individual submission

- Submit the zip file `COM6005_pXXXXXX_Report.zip` to Moodle, where `pXXXXXX` is your student ID number.

- The zip file **MUST** contain:
  - One folder `COM6005_pXXXXXX_Code`
    - to contain all data, and Python program files (in jupyter notebook `.ipynb` format)
  - Files:
    - `COM6005_pXXXXXX_WrittenReport.pdf`
    - `COM6005_pXXXXXX_Veriguide_Report.pdf`
    - `COM6005_pXXXXXX_Report_CoverSheet.pdf` (**Annex 2**)

Item 3 - Deadline: 06 Dec 2025 (Sat) 23:59 (Week 14) – individual submission

- Submit the Peer Evaluation Form, with the convention: `COM6005_L0Z_GYY_PeerEvaluation_pXXXXXX.pdf`

Penalty (for all project assessment components):

- For late submission, zero marks will be given to the respective project component.
- For plagiarism: zero marks will be given to the whole project (As such, students will receive a failure grade for the module). The case will also be reported to the University and other disciplinary actions will be taken by the University.

- End of Project Guideline -

Annex 1

The Hang Seng University of Hong Kong

Department of Computer Science

COM6005 Statistical Modelling (2025-2026 Sem 1)

Group Presentation Cover Sheet (30%)

Group Number	**L0* _G****	**Student ID**	**p**	**Student Name**	
Presentation Date/Time					
Domain Studied					

	Assessment Criteria	*MARKER USE ONLY*
Contents	Knowledge of the Subject Matter	___ out of 5
	Organization and Accuracy	___ out of 10
	Value of insights	___ out of 5
Presentation	Proficiency of Language	___ out of 5
	Communication Skills	___ out of 3
	Response to Audiences' Comments	___ out of 2
	Total:	___ out of 30

Deadline: **22 NOV, 2025 (SAT), 23:59**

Annex 2

The Hang Seng University of Hong Kong

Department of Computer Science

COM6005 Statistical Modelling (2025-2026 Sem 1)

Written Report Cover Sheet (55%)

Group Number	**L0* _G****	**Student ID**	**p**	**Student Name**	
Presentation Date/Time					
Domain Studied					

	Assessment Criteria	Marker use only
Contents	Introduction	___ out of 5
	Methodology and Results	___ out of 10
	Discussion and Conclusion	___ out of 15
Organization and Style	Proficiency of Language	___ out of 5
	Structure and Organization	___ out of 5
Computing Techniques	Correctness of Program	___ out of 10
	Program Effectiveness	___ out of 2
	Appropriate use of statistical methods, visualization and tables	___ out of 3
	Total:	___ out of 55

Deadline: **6 DEC, 2025 (SAT), 23:59**

Annex 3

The Hang Seng University of Hong Kong

Department of Computer Science

COM6004 Data Mining (2025-2026 Sem 1)

Peer Evaluation Form (15%)

Group Number	**L0* _G****	**Student ID**	**p**	**Student Name**	
Presentation Date/Time					
Domain Studied					
Please indicate your contribution to the project below					

Other Group Members		*Comments*	**Mark by student**
Student ID	*Student Name*		
			___ out of 10
			___ out of 10
			___ out of 10
			___ out of 10

Assessment Criteria	**Mark by instructor only**
Comments on other groups	___ out of 5

Deadline: **06 DEC, 2025 (SAT), 23:59**

Assessment Criteria for COM6005 Project

A. Presentation (in English)	30%
1. Knowledge of the Subject Matter (5%)	
2. Organization and Accuracy (10%)	
3. Value of insights (5%)	
4. Proficiency of Language (5%)	
5. Communication Skills (3%)	
6. Response to Audiences' Comments (2%)	
B. Individual Written Report	55%
1. Introduction (5%)	
2. Methodology and Results (10%)	
3. Discussion and Conclusion (15%)	
4. Proficiency of Language (5%)	
5. Structure and Organization (5%)	
6. Correctness of Program (10%)	
7. Program Effectiveness (2%)	
8. Appropriate use of statistical methods, visualization and tables (3%)	
C. Peer Evaluation Form	15%
1. Contribution & Peer Marks (10%)	
2. Comments on other groups (5%)	

A1. Knowledge of the Subject Matter (5%)

Mark range	Expectations
5	Demonstrates thorough understanding of the topic and the full statistical modeling and data mining lifecycle. Clearly explains the rationale for chosen modeling techniques and methodology using Python and justifies their suitability for addressing the societal problem aligned with SDGs.
4	Shows solid grasp of key concepts and techniques. Explains what was done and why, but with limited discussion of limitations, alternatives, or deeper rationale behind methodological choices.
3-2	Demonstrates basic understanding. Can describe the steps taken but struggles to articulate the underlying principles or justify selection of methods and tools in detail.
1-0	Displays significant knowledge gaps. Cannot explain key concepts, methods, or project steps clearly. Provides inaccurate or irrelevant information.

A2. Organization and Accuracy (10%)

Mark range	Expectations
10-8	Presentation is exceptionally well-structured. The flow from introduction to conclusion is seamless and compelling. All information, data, results, and visualizations are accurate and meticulously support the narrative of building the index.
7-5	Presentation is clearly organized with a logical sequence. Information is generally accurate. Slides are clear and support the talk, though some transitions may be slightly awkward or some details may be slightly misrepresented.
4-2	The presentation has a basic structure but may be jumpy or illogical. Contains some inaccuracies or inconsistencies in data presentation. Slides may be overly text-heavy or confusing.

1-0	Disorganized and difficult to follow. Lacks a clear story. Information is predominantly inaccurate, poorly sourced, or presented incoherently.

A3. Value of insights (5%)

Mark range	Expectations
5	Offers original, critical, and insightful analysis. Clearly explains the "so what" of the findings for urban planning/public health. Draws meaningful, non-obvious conclusions from the "Well-Being Index" and provides actionable, data-driven recommendations.
4	Provides clear analysis and some good insights. Explain the findings and their implications logically. Conclusions are well-supported but may be more predictable or less actionable.
3-2	Insights are mostly descriptive (e.g., "this neighborhood scored higher"). Restates findings without deep interpretation or connection to the project's mission. Conclusions are obvious or simplistic.
1-0	Lacks any meaningful analysis. Merely lists results or reads data from slides. Fails to draw any valid conclusions or provide insights.

A4. Proficiency of Language (5%)

Mark range	Expectations
5	Uses a wide range of vocabulary and complex sentence structures accurately. Grammar and pronunciation are virtually error-free, enhancing clarity and professional tone. Technical terms are used with precision.
4	Language is clear and effective with minor errors that do not impede understanding. Pronunciation is generally clear. Uses technical terms correctly most of the time.
3-2	Language is simplistic and contains frequent grammatical errors that occasionally obscure meaning. Pronunciation may be difficult to understand, affecting clarity.
1-0	Language is incoherent or filled with severe errors that make the presentation very difficult to follow. Pronunciation is poor.

A5. Communication Skills (3%)

Mark range	Expectations
3	Engages the audience masterfully through confident eye contact, clear and varied pacing, and expressive tone. Uses slides as a visual aid, not a script. Well-rehearsed and polished. Stays within the 20-minute limit.
2	Clear and competent delivery. Maintains good eye contact and a steady pace. Uses visual aids effectively. May have minor issues with posture, over-reliance on notes, or time management.
1	Delivery is hesitant and uneven. Reads heavily from slides or notes. Little eye contact. Pace may be too fast or slow. Presentation may be significantly under or over time.
0	Communication is poor. Mumbling, monotone, no engagement with the audience. Completely reliant on reading text.

A6. Response to Audiences' Comments (2%)

Mark range	Expectations
2	Listens carefully to questions and responds thoughtfully, confidently, and accurately. Elaborate on answers with deeper insights, clearly demonstrating a strong understanding. Handles challenging questions with poise.

1	Answers might be brief or somewhat simplistic. May need a moment to formulate a response or ask for a question to be repeated.
0	Unable to answer questions effectively. Responses may be incorrect, vague, unrelated to the question, or indicate a significant gap in understanding. May be defensive, or simply state "I don't know" without attempting to engage with the question's topic.

B1. Introduction (5%)

Mark range	Expectations
5	Introduction is engaging and clearly sets the context. The problem is explicitly defined within the framework of the chosen societal issue and its connection to an SDG. Background is detailed and relevant. Objectives are specific, measurable, and clearly articulate the value of the study.
4	Introduction is clear and logically structured. Problem, background, and objectives are defined and relevant, though may lack some depth or precision.
3-2	Introduction covers the basics but is too general or vague. Problem statement may be unclear, and objectives are not well-defined or measurable.
1-0	Introduction is poorly constructed, missing key elements, or fails to explain the purpose and significance of the project.

B2. Methodology and Results (10%)

Mark range	Expectations
10-8	Methodology is thoroughly detailed and reproducible, covering data cleaning, preprocessing, algorithms with parameters, and evaluation metrics. Results are presented with clear, professional visualizations (charts, tables) that are well-labeled and thoughtfully interpreted.
7-5	Methodology is clearly described with minor missing details (e.g., parameter settings or library versions). Results are presented effectively with appropriate visualizations, though some explanations could be clearer.
4-2	Methodology lacks sufficient detail, making replication difficult. Results are shown, but visualizations may be poorly chosen, unlabeled, or lack proper interpretation.
1-0	Methodology is unclear or missing. Results are absent, incorrect, or incomprehensible.

B3. Discussion and Conclusion (15%)

Mark range	Expectations
15-12	Provides insightful interpretation of results, clearly connecting findings to objectives and societal context (e.g., SDGs). Thoughtfully discusses limitations and their impact. Conclusions offer powerful, actionable recommendations for stakeholders.
11-7	Logically interprets results with connection to objectives. Conclusions summarize key findings and relevant recommendations, but may lack depth in critical evaluation or limitation discussion.
6-2	Mainly restates results with minimal interpretation. Conclusions are simplistic, listing obvious points without deeper insights or actionable guidance.
1-0	Discussion and conclusion are unclear, missing, or unrelated to results. Fails to provide meaningful interpretation or recommendations.

B4. Proficiency of Language (5%)

Mark range	Expectations

5	The report is written in a clear, concise, and scholarly manner. Flawless grammar, spelling, punctuation, and word choice. Tone is professional and appropriate for an academic/technical report.
4	The report is clearly written with minor errors in grammar or word choice that do not impede understanding.
3-2	The report contains frequent errors in language that occasionally obscure the meaning. Writing style is simplistic or awkward.
1-0	The report is poorly written with severe errors that make it very difficult to understand.

B5. Structure and Organization (5%)

Mark range	Expectations
5	The report is perfectly structured according to guidelines. Formatting (font, spacing, margins) is strictly adhered to. The logical flow between sections is seamless. Writing is concise and remains within the 10-page limit.
4	The report is well-structured and formatted correctly. The flow is logical. May be slightly over or under the page limit or have minor formatting inconsistencies.
3-2	The report has a structure, but it is awkward in places. Formatting has several errors (e.g., wrong spacing, font). The flow is sometimes hard to follow.
1-0	The report is disorganized and difficult to follow. Fails to meet basic formatting requirements.

B6. Correctness of Program (10%)

Mark range	Expectations
10-8	Code is fully functional, well-documented (with comments), and runs without errors. It correctly implements all the described data mining processes and produces the results shown in the report. Code is clean and efficiently structured.
7-5	Code is functional and produces correct results. Documentation may be sparse in places, or the code may have minor inefficiencies, but it is generally correct and runnable.
4-2	Code runs but may have bugs that require fixing to produce the correct results. Documentation is poor. Code is messy or inefficient.
1-0	Code does not run, is completely incorrect, is missing, or is plagiarized.

B7. Program Effectiveness (2%)

Mark range	Expectations
2	The code is not only correct but also efficient and well-structured.
1-0	The code, while it may produce the correct output, is poorly designed or inefficient.

B8. Appropriate use of statistical methods, visualization and tables (3%)

Mark range	Expectations
3	Flawless, insightful choice and execution of methods and visuals. Professional, publication-quality.
2	Generally appropriate and correct. Functional and clear but not optimal.
1-0	Poor, inappropriate, or misleading choices. Detracts from the report.

C1. Contribution & Peer Marks (10%)

Mark range	Expectations
10-8	Described by peers as an outstanding contributor who went above and beyond. Took leadership, completed a large share of high-quality work, was reliable, and helped other group members. Peer marks are consistently high.
7-5	Described as a solid, reliable contributor who completed their assigned tasks on time and to a good standard. Collaborated well with the team. Peer marks are good.
4-2	Described as contributing minimally or inconsistently. Work may have been late, low quality, or required significant revision by others. Peer marks are average or low.
1-0	Described as contributing little to nothing, being unresponsive, or hindering the group's progress. Peer marks are very low. May trigger a meeting with the instructor.

C2. Comments on Other Groups (5%)

Mark range	Expectations
5	Questions or comments for other groups are insightful, constructive, and demonstrate careful listening and critical thinking about the data mining process and findings.
4	Questions are relevant and show understanding of the presented material. They may be more clarificatory than deeply insightful.
3-2	Questions are very basic, off-topic, or suggest a lack of attention during the presentation.
1-0	No questions asked, or questions are inappropriate or disruptive.