STAT4011 – Project 2: Statistical Data Analysis

Instructor: Dr. HO, Kwok Wah (LSB116, remus@cuhk.edu.hk)

Purpose of the Project:

- Apply the statistical techniques that you have learnt to analyze a sizable data set.
- Gain experience in developing statistical models, performing data analysis, performing group presentation and writing a report.

Background:

The datasets given are reported crime record data of Los Angeles (LA) from 2010 to 2019 and 2020 to 2025. These real datasets include time, location, crime type, victim characteristics and other variables for each reported case. Imagine that you are a data analyst working for the LA police department (LAPD). You would like to see how the data can help LAPD understand more about the crime patterns and trends in the city. You also want to see if you can provide suggestions based on data to help fighting against crime in the city. Since the data are real, you can (but not necessary) supplement with other public real datasets that you find helpful to enrich your analysis. The provided data files are very large and you are not expected to be comprehensive. You can solely focus on a part of the datasets that can help on the research questions you are interested in.

Data files:

There are 2 data files of similar characteristics uploaded on Blackboard for you to download. One for the reported crime records from 2010 to 2019 (downloaded at 29/9/2025). The other for reported crime records from 2020 to 2025 (downloaded at 29/9/2025). They are downloaded from:

https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data
https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

Variable descriptions and other information can be found in the webpages of the above links.

Some guidelines for your analysis:

1. The first step for any data analysis is to get familiarized with the data (the part of the datasets that you focus on). Data cleaning may be needed for missing data, outliers and possible recording errors. Exploratory data analysis to discover any special feature of the variables or even the relationships between the variables. Using simple descriptive statistics, graphical or numerical, to understand the distributions of the variables.
2. The next step is to write down research questions that you think are interesting and can possibly be answered by the data you use. Usually the preliminary work in the first step can shed light on the direction and limitations of your study.
3. Try to answer your research questions using any statistical methods that you have learnt (e.g. regression, time series models, testing, Bayesian analysis, or even machine learning methods). You can use any statistical software.
4. Lastly summarize your findings. Hopefully you can provide some insightful comments to LAPD based on your findings.

Group formation:

You are allowed to form groups of 4 to 6 students. **Please send me through email your group list on or before 17th Oct.** If you are not formed into groups by that time, I will randomly assign you into a group.

Remark: for groups with fewer than 5 students, there is a chance that I will add student(s) into your group.

Consultation:

1. **On 21st Oct and 11th Nov, I will be in SC L2 from 4:30pm to 5:15pm.** If you have questions about the project, you can come to ask me. Of course, you can also email me any other time when you have questions about the project.
2. I will schedule a separate zoom meeting with each group on **28th Oct** or **4th Nov**. Schedule for each group will be confirmed later. Your group should briefly tell me your plan and progress in the meeting. I will also give some advice to you as well.

Group Presentation: All group members have to speak in the presentation and the presentation should be conducted in English. 15 minutes for each group and I will ask questions or give comments after each presentation. **Presentation for each group will be on 18th Nov 2025 or 25th Nov 2025 during lecture time in SC L2.** As all groups are basically working on the same data, you cannot listen to presentations of other groups. We shall confirm about the presentation schedules later.

Group Report: Each group should submit one report. The report MUST be typed and at most 20 pages (A4 sized, Font size 12) including everything. The report should be well organized that shows clearly your methods, analysis and results. **Please submit the report through VeriGuide, referred as Project 2, on or before 9th Dec 2025**. Your group should assign one member to **upload on Blackboard your code (and extra data if any) on or before 9th Dec 2025 .** Students should pay attention to the academic honesty and plagiarism policy of the University. Marks may be deducted for hints of plagiarism reported by VeriGuide. Case with sufficient evidence of plagiarism will be forwarded to the Disciplinary Committee of the Science Faculty. Regarding use of AI policy, we adopt approach 3 that the use of AI tools is allowed with explicit acknowledgement. Students are required to acknowledge all functional uses of an AI tool and cite it when they paraphrase, quote, or incorporate into their own work any content that was created by it. For example, acknowledgement can be written like "I acknowledge the use of – e.g. ChatGPT (Mar 20 version) to, e.g. plan my report, generate some ideas for the content, etc."

Grading System:

Report accounts for 28%

Remark: the report will be assessed based on innovation, level of difficulty, accuracy of computation and clarity of report presentation. References for any idea or method from external sources should be cited at the end of the report.

Presentation accounts for 22%

Remark: while mainly based on the performance of the whole group, presentation mark for each student will be tuned according to individual performance during presentation.

Each student MUST upload on Blackboard his/her own assessment on the percentage of contribution of each team member in the project on or before 9th Dec 2025. You should try to be as objective as possible when judging the percentage of contribution. Follow-up meetings may be arranged (if needed) for groups with very uneven reported contributions to understand the need for mark adjustments of some group members.