IND5003 Project Information

Contents

1 Project Titles

1. The python notebook `nea_radar_images.ipynb` contains code that will download images of rain areas surrounding singapore, similar to Figure 1.

   Utilise computer vision techniques to track and study the trajectories of storms in the region.

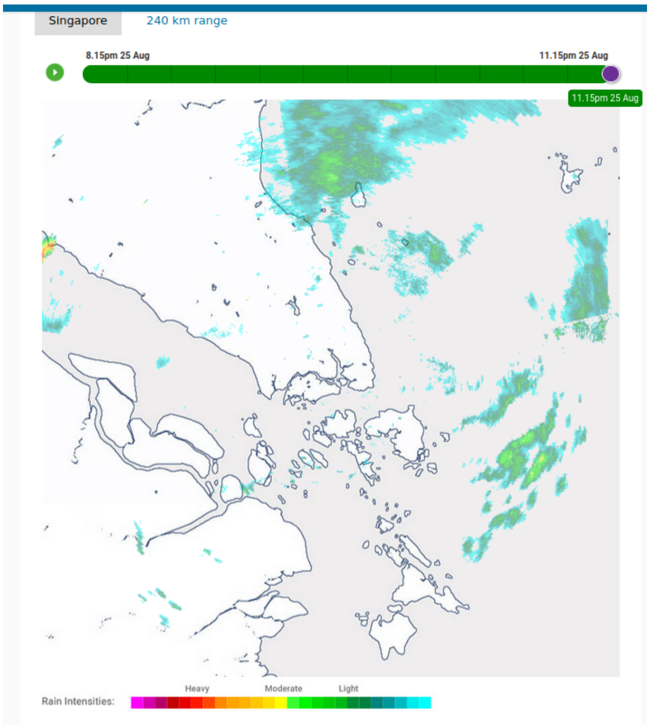

Figure 1: Rain areas

2. The python notebook `lta_images.ipynb` contains code that will download images of traffic at various major roads in Singapore, similar to Figure 2.

   Utilise computer vision techniques to study the traffic jams at the causeway and second link.

Figure 2: Causeway (left: towards Johor, right: towards BKE)

ℹ Note

I am aware that our topic on computer vision comes towards the end of the course, but I intend to put the notes up during recess week at the latest so that you can read up.
In the meantime, if you choose to work on one of the above projects, I strongly encourage you to sign up and complete the free bootcamp by the opencv authors. It is very useful indeed!

3. The World Values Survey is conducted annually in countries around the world. The survey aims to track and study the values and norms of residents of various countries. The website contains data, documentation and reports for previous years. For this project, our coursemates will be asked to fill up a modified version of the survey (compulsory). Your task is to carry out an analysis of this data from our coursemates.

4. The Enron Email corpus consists of emails from 150 users, mostly senior management. It was made public during an investigation into the Enron scandal. Your task is to carry out an unsupervised learning analysis of the Enron corpus.

5. Implement an agent-based model of interest to you. Here are some possibilities:

   - Penalty-takers in Soccer
   - Maritime pirate attacks
   - Negotiation in e-commerce

   The papers above contain detailed descriptions of agent-based models. Your task is to implement *basic* versions of them using mesa, and then to analyse various scenarios. Parametrising simulation models requires real data, so part of the task here is to decide what parameters to decide for the model.

ℹ Note

Again, I will put up the notes for simulation models very soon.

6. Self-proposed project. If you have a problem from your workplace, or you prefer to work on a different from one of the above, that is perfectly fine too. However, projects need to contain certain characteristics:
   - The data should be prepared and cleaned carefully. With the tools we have today, the *analysis* burden is less than before. Hence as analysts we can pay more attention to data preparation. This phase should include looking for multiple datasets to allow for checking your findings against, and datasets of additional but relevant features to improve your analysis. If not sure, do check with me.
   - The project should require you to apply one or more of the topics in our course, or to learn a new method on your own.
   - The project cannot consist of only supervised learning.
   - The research question should be clearly defined. It cannot be identical to analyses performed online.

Projects need to be finalised before the end of week 5. If you are going to propose your own project, please send me a 1-page write-up with your proposed idea before Friday of week 4.

2 Submissions

The following are the submissions for the group project:

1. Video presentation
   - A 10-minute video to be uploaded to a sharepoint folder (to be created closer to the submission date). The presentation should communicate the major findings from your data analysis. The portion on methods review can be kept to a minimum.
   - You can assume that the audience for the presentation is a technical audience, i.e. it is a knowledge-sharing presentation.
2. Report
   - A data analysis report submitted as a pdf document.
   - The maximum number of pages should be 20, excluding cover, contents, references and appendices. The font size cannot be smaller than 10 Times New Roman.
3. Source codes
   - A zip file containing all the Python source codes written for the project.
   - Codes can be in the form of python notebooks, or in the form of scripts/modules, or a combination of both.

The deadline for all the submissions is 2359hrs on 3 November 2025.

3 Grading (50%)

The total marks for the project is 50% of your final grade. This is broken down into 40% for a common group grade, and 10% for an individual grade.

3.1 Group grade (Total: 40%)

Video presentation (Subtotal: 10%)

The video presentation will be graded on clarity of content and creativity.

Data analysis report (Subtotal: 20%)

You are encouraged to include the following sections in your report (other sections are up to your discretion):

1. Methods review. This section could be a review of relevant/existing work similar to the task, or it could be a quick introduction to the model used in the report (if it is not one that was covered in the lecture notes).
2. Statement of questions of interest. These are the 2 - 3 research questions that you aim to answer with this analysis.
3. Data overview. This section is where you can describe the data collection and preparation procedures. Credit will be given for finding relevant datasets that provide additional evidence/context for your conclusions.
4. Overview of code structure.

The data analysis report will be graded on the quality of the analysis and the clarity of the report. In terms of clarity of report, we look for the use of proper English (no bullet points). We also encourage the use of visualisations. The report should be well-organised.

In terms of quality of analysis, we look for iteration in the analysis, since every analysis reveals a bit more about the data, and suggests what could be done next. The findings pertaining to the questions of interest should be communicated clearly, demonstrating understanding of techniques. A discussion of the limitations of the analysis should also be present. Evidence of self-learning, where you identify a new technique, or go deeper into one of the techniques introduced in the lectures) also counts as a plus point.

A very insightful paper by prominent statisticians (McGowan, Peng, and Hicks, 2023), breaks down every data analysis into six areas that can be used to assess its quality. These six areas are: data quality, exhaustiveness, skepticism, second-order questions, clarity and reproducibilty of analysis. You may want to skim through the paper when assessing the quality of your own report.

Please note that NUS has a policy on the use of AI for homework. I encourage you to use AI models to assist with writing code, but do ensure that you do not violate the university guidelines. Especially when writing reports, bear in mind that these models do tend to invent references, authors and titles.

Code (Subtotal: 10%)

Code is assessed in terms of reproducibility, documentation (via comments/docstrings/README files), and organisation of folders.

3.2 Individual grade (10%)

The individual component will kick in after the submissions. The score comes from:

1. Your rating by fellow group members.
2. Individualised quiz questions on your own project. This quiz will be conducted in the week 13 lecture time. It may include asking about the source codes that your group submitted.

3.3 Acdaemic references

McGowan, Lucy D'Agostino, Roger D Peng, and Stephanie C Hick, 2023. "Design Principles for Data Analysis," *Journal of Computational and Grpahical Statistics*, **32** (2): 754–61.