Instructions, Deliverables & Naming Conventions:

In this project, you are given a dataset collected by an actual IoT system (see description below) and asked to use the dataset to build a forecasting model. You have to answer a set of questions (there are no fixed answers), as well as propose your own interesting questions.

(a) Form teams in groups of 4 students and register your group under NUS Canvas → EE4211/TEE4211 → People → Group. Take note of your group number.
(b) For each of the 3 sub-parts below, submit a single zip file containing (i) PDF file/Print preview of your iPython notebook, (ii) the original iPython notebook with all your code (ipynb file), (iii) any additional data files required to run the notebook. Use markdown cells in the ipynb file to elaborate and provide your answers to the questions.
    (i) Question 1: Data Cleaning & Exploring the Data (10 marks). Please name your zip file {Group Number}_Question_1.zip (e.g., If your team is group 42, 42_Question1.zip) and upload to Canvas in the correct assignment by the deadline.
    (ii) Question 2: Forecasting (10 marks). Please name your zip file {Group Number}_Question2.zip and upload to Canvas in the correct assignment by the deadline.
    (iii) Question 3: Group Proposed Project (10 marks). Please name your zip file {Group Number}_Question3.zip and upload to Canvas in the correct assignment by the deadline.
(c) Presentation and Report (10 marks): Prepare slides, report and a video presentation regarding your response to Question 3.
    (i) Zip the slides and report and name this zip file {Group Number}_SlidesAndReport.zip and upload to Canvas in the correct assignment by the deadline.
    (ii) Name your video file {Group Number}_Presentation.mp4 and upload to Canvas in the correct assignment by the deadline.
(d) In summary, the project carries a total of 40 marks. There are 4 deliverables: Question 1 including group project proposal (10 marks), Question 2 (10 marks), Question 3 (10 marks), and Presentation and Report (10 marks).

Target Data (Predicted/Output/Response Variable) Description:

In this project, we will consider the carpark availability dataset provided by the Singapore government at this link [1] for use as target data (predicted/output/response variable). Supplementary information of the carparks are provided at this link [2].

An example of using Python to make a data.gov.sg API call for a single time instance is shown in the provided sample code: "ExampleAPI.ipynb". **You will have to modify the provided sample code (or write your own code)** to collate data from multiple time instances together.

Note that the data.gov.sg API returns the data as a JSON (JavaScript Object Notation) object. The provided sample code transforms this JSON object into a pandas dataframe. An example of the data from the provided sample code is shown below:

	carpark_number	update_datetime	total_lots	lot_type	lots_available
0	HE12	2022-04-12T12:12:32	105	C	0
1	HLM	2022-04-12T12:12:42	583	C	0
2	RHM	2022-04-12T12:12:32	329	C	106

[1] https://data.gov.sg/datasets/d_ca933a644e55d34fe21f28b8052fac63/view
[2] https://data.gov.sg/datasets/d_23f946fa557947f93a8043bbef41dd09/view

Questions:

1. Data Cleaning & Exploring the Data (10 marks)

1.1 Look at the features in the dataset. What do the values in the column "lot_type" mean? Hint: Note that data.gov.sg gets its data from the Land Transport Authority (LTA) too. Try searching for the LTA Datamall API documentation.

1.2 Carry out and document a systematic approach to approximate the frequency at which the data values are updated. Note: The purpose of this question is to avoid querying for data unnecessarily. Although the API date_time parameter is specified to seconds, the database may not be updated every second.

1.3 (i) How many unique carparks are included in the carpark availability dataset? (ii) Check if this value varies with time. Explain why this check is important (about 20 words).

1.4 A carpark may have malfunctioning sensors. There are many types of possible malfunctions. Identify one of these carparks that you believe has a malfunctioning sensors. Explain what the "malfunction" is in this case (about 20 words).

1.5 Create a dataset of hourly carpark availability **(i.e., for this project, use the ratio: lots_available/total_lots)** from the raw data for the month of July 2024. Plot the average (average across all carparks) hourly carpark availability against time for that interval. Identify any patterns in the plot (about 50 words).

Note: You will have to decide what to do if there are no carpark readings for a certain hour. For example, some may impute the missing data or ignore it. You also have to decide if you want to (i) compute the ratios for each carpark and then average OR (ii) compute the total lots_available and total total_lots and take the ratio.

1.6 Intuitively, we expect that carpark availability across certain carparks to be correlated. For example, many housing carparks would experience higher carpark availability during working hours. Using the same interval chosen in 1.5, pick a carpark and find the carpark that is most correlated to it (in terms of carpark availability). State the type of correlation used (e.g. Spearman, Pearson, etc).

1.7 Group Project Proposal for Question 3: Please include a short proposal (around 500 words) of what your team intends to do for the Group Proposed Project in Question 3. For the group project proposal, you may use additional datasets to supplement your analysis or look at unaggregated data, etc. See Question 3 below for more information about this. Please use markdown in the iPython notebook to present your proposal.

2. Forecasting (10 marks)

2.1 In this part, you will build a model to forecast the hourly carpark availability in the future (averaged across all carparks instead of looking at each carpark individually). Can you explain why you may want to forecast the carpark availability in the future? Who would find this information valuable? What can you do if you have a good forecasting model?

2.2 Build a *ridge regression* model to forecast the hourly carpark availability for a given month. Use the month of July 2024 as a training dataset and the month of August 2024 as the test dataset. For this part, do not use additional datasets. The target is the hourly carpark availability and you will have to decide what features you want to use. Generate two plots: (i) Time series plot of the actual and predicted hourly values (ii) Scatter plot of actual vs predicted hourly values (along with a line showing how good the fit is).

2.3 Do the same as Question 2.2 above but use random forest (RF) regressor. Be sure to state clearly all parameters of your chosen model.

2.4 Do the same as Question 2.2 above but use multi-layer perceptron (MLP). Be sure to state clearly all parameters of your chosen model.

2.5 Make a final recommendation for the best regression model (out of the 3 methods above) by choosing suitable performance metric(s). To ensure a fair comparison, carry out hyperparameter tuning for all 3 methods. Then, make a final recommendation selecting only one model. Include both quantitative and qualitative arguments for your choice.

3. Group Proposed Project (10 marks)

3.1 At this point, you understand the data well. For your group proposed project, you must explore some aspects of machine learning models. You must use the dataset given but you may use additional datasets to supplement your analysis (e.g., weather data), look at unaggregated data, look at the difference in carpark availability for carparks with free parking, etc. Note that you are not limited to the initial proposal and are free to expand on it.

3.2 Based on the insights derived from the analysis, suggest a practical action that can be taken in approximately 200 words (i.e., an action that can be taken to benefit society. Do not suggest actions such as hyperparameter tuning here). You do not need to carry out the action. A simple example is, if the team had made models to predict carpark availability of individual carparks in Q3.1, then these models can be used to develop an application to the public that forecasts carpark availability to reduce congestion in carparks during peak hours.

4. Presentation and Report (10 marks)

4.1 Prepare slides, report and a video presentation regarding your group's contribution to Question 3, the group proposed project. The presentation and report should cover the analysis done in Question 3: Group Proposed Project. Note: Do not cover Questions 1 and 2 in the presentation and report.

4.2 Slides and report: Limit the number of slides to 15 and limit the report to 10 pages. Please include the following points at a minimum with equal emphasis on all points.

i. Problem Description – What is the problem that you are solving, what is the context, etc. Why your proposal is an important problem worth working on.

ii. Experimental Setup – Describe any assumptions you make, experiment setup , etc.

iii. Results and Analysis – Visualize your results in a clear and easy to understand manner. What insights do you get from your results, key reflections, etc.

iv. Reflections & Conclusions – What practical benefits does your work bring? How can your work be used to benefit society? What is the takeaway message, what did you learn, etc.

4.3 Video: Make a 10-12 minute video for your group's presentation. Each group member must present in the video. Please convert your video to mp4 format with a minimum resolution of 480p.