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.

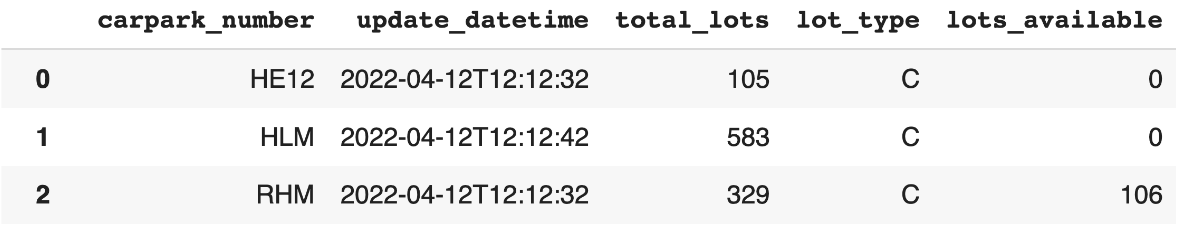
1. Form teams in groups of 4 students and tegister your group under NUS Luminus → EE4211/TEE4211 → Class & Groups → Class Groups. Take note of your group number. If you face any difficulties, please contact the teaching team.
2. For each of the 3 sub-parts below, use a iPython notebook notebook (ipynb file) to do the analysis and answer all the parts of the Question. **Use markdown in the ipynb file itself to elaborate and provide your answers to the questions.** The iPython notebook should form your report (i.e., your report should not be a separate document file).
3. For each of the 3 sub-parts below, submit a single zip file containing (i) PDF file/Print preview of your Jupyter notebook, (ii) the original Jupyter notebook with all your code (ipynb file), (iii) any additional data files required to run the notebook.
   (a) Complete Question 1. Please name your zip file Group_Number_Question_1.zip (e.g., EE4211_Group_1_Question_1.zip) and upload to LumiNUS by the appropriate deadline.
   (b) Complete Question 2. Please name your zip file Group_Number_Question_2.zip and upload to LumiNUS by the appropriate deadline.
   (c) Complete Question 3. Please name your zip file Group_Number_Question_3.zip and upload to LumiNUS by the appropriate deadline.
4. Prepare slides and a video presentation regarding your response to Question 3. Please zip your slides and video file together, name your zip file GroupName_Presentation.zip and upload to LumiNUS by the appropriate deadline.
5. 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 (10 marks).

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

In this project, we will consider the carpark availability dataset provided by the Singapore government's data.gov.sg (`https://data.gov.sg/dataset/carpark-availability`) for use as target data (predicted/output/response variable). Supplementary information of the carparks are provided at `https://data.gov.sg/dataset/hdb-carpark-information`.

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: "EE4211-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

Questions:

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

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

1.2 Try making an API call for the data from a single specified date & time. Then, do the same thing for the next second of the initially chosen date & time. Notice that "update_time" is unchanged. 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 carparks are included in the data.gov.sg car park database? (ii) Does this number vary based on the time? You should notice that it does vary with time. (iii) A carpark may have malfunctioning sensors and nor report its data. Identify one of these carparks with anomalous sensors and a time period where that carpark's sensors were malfunctioning.

1.4 Generate hourly readings from the raw data. Select a one month interval and plot the hourly data (time-series) for that interval (aggregate results instead of plotting for each location individually). Identify any patterns in the visualization. Note: You will have to decide what to do if there are no carpark readings for a certain hour, for example, should you impute the missing data or ignore it.

1.5 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 1.4, write a function to find the top five carparks with which it shows the highest correlation. Demonstrate an example of this function call using a randomly selected carpark.

1.6 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 (aggregated 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 linear regression model to forecast the hourly carpark availability for a given month. Use the month of July 2022 as a training dataset and the month of August 2022 as the test dataset. For this part, do not use additional datasets. The target is the hourly carpark availability percentage 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 support vector regressor (SVR).

2.4 Do the same as Question 2.2 above but use decision tree (DT) regressor.

2.5 Make a final recommendation for the best regression model (out of the 3 methods above) by choosing a suitable performance metric. 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 quite well. Carry out the analysis you proposed in your group project proposal. You should use the dataset given but you may also use additional datasets to supplement your analysis, look at unaggregated data, etc. Please be sure to justify why the analysis is useful and interesting in the context of a data science project. 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 (i.e., an action that can be taken to benefit society. Do not suggest actions such as hyperparameter tuning here).

4. Presentation (10 marks)

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

4.2 Slides: Limit the number of slides to 15 slides maximum.

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.

For your benefit, here are some pointers from former students in the class:

1. The project can be an opportunity to talk about data science in interviews, use the project wisely.
2. Keep the explanations concise and use subplots to compress the plots if there are too many plots.
3. To increase the reproducibility of results, Google Collab is a good option (which allows collaboration too).
4. Showing the version of python packages that are used can increase the reproducibility score of the project.
5. Setting random seeds when using randomized operations like train_test_split will also help with reproducing the results.
6. Start the projects early because it is too much work to rush last-minute.
7. Do the projects yourself, experiment and understand because you will learn a lot.