School Name	School of Computing
Semester	AY2022-2023 Semester 2
Course Name	DAAA
Module Code	ST1508
Module Name	Practical AI

Background

- Just Taxi is a ride-hailing service company. The company hired your team as their data science team to develop an easy to use, and intuitive machine learning software application that may help them to visualize, and analyze **taxi** data.

- They want the application to be portable, and able to run on the manager’s laptop. Therefore, during the company meeting, if questions arises concerning driving safety, the application can then be used to help to provide answers.

- The management of Just Taxi also want the data science team to build an interactive dashboard for data visualization, which could provide insights in the bigger picture of the ride-hailing operation. For instance, what kind of driving behavior will lead to danger?

Data Source

The company has multiple historical datasets stored in *.csv format. It contains ~7 million **taxi safety data**.

CA1 (100%)

The whole project will be split into 2 phases. The objective of CA1 is to help you gain a better understanding of the data science project workflow, including creating SQL database for data storage, setting up data pipeline between SQL and Python, using Python for data analysis and building interactive dashboard for business users.

Guidelines

1. You are going to work on the data science project on a group basis (3 students per group).
2. You are going to work on the data science project on a group basis (3 students per group).
3. In CA1, you will be working on the first part of a data science project and write a report that describes your solution to the tasks.
4. Write a Jupyter Notebook including your code, comments, and data visualization.
5. Create a nice presentation slide for your project.
6. Students are required to submit their assignment using the assignment link under the CA1 folder. Please remember to include student names and student numbers in the notebook and slides (using zip file, name of file, a student name+class, e.g. John-2B02)
7. The normal SP's academic policies on Copyright and Plagiarism applies. Please note that you are to cite all sources. You may refer to the citation guide available at: http://eliser.lib.sp.edu.sg/elsr_website/Html/citation.pdf

Submission Details:

Deadline: 2022-12-02 23:59H
Submit through BrightSpace

Late Submission
50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter. Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

Task:

1. **Build a SQL database** for the taxi data storage. You are required to use Microsoft SQL Server. In order to test the functionality of the SQL database, you are also required to **write 3 complex SQL queries** to extract useful insights from the data.

2. **Create an ETL pipeline** between SQL database and Python, in order to facilitate the later data analysis. You can use Python libraries to build the ETL pipeline, or you can also use more advanced third-party ETL tools to achieve the purpose.

3. After loading the data from SQL database to Python, you are required to write Python code in your **Jupyter Notebook** to perform data pre-processing, data cleansing, exploratory data analysis and generate some data visualization using advanced libraries (e.g. Seaborn, Bokeh).

4. You are required to **create an interactive dashboard**. You can use Tableau or Power BI to build the dashboard.

CA1 deliverables:

1. A well-written **project report** (word document, < 20 pages), describing all your work done in phase one.

2. A **Jupyter Notebook** (*.ipynb)**,** including data pre-processing, data cleansing, exploratory data analysis and simple data visualization.

3. A **Tableau** or **Power BI dashboard** file (*.twbx or *.pbix)**.**

4. **PowerPoint Slides** (< 30 pages) for a 20-minute presentation.

5. **Scrum Document**: User Story, Product Backlog, Bi-Weekly Scrum Reports.

6. **Peer Evaluation Form** (every individual student needs to submit the form).

CA1 Evaluation Criteria:

SQL database creation, ERD, SQL queries & ETL pipeline set up	15 %
Data cleansing and wrangling	10 %
Exploratory data analysis in Python	10 %
Building Dashboard	30 %
Phase-1 Presentation & Slides	10 %
Phase-1 Report & Scrum Documentation	15 %
Technical Complexity	10 %

-- END –

Appendix 1: User Stories

Users

Based on the team's preliminary discussions with the company, it has been concluded that they need to address the needs of the following groups of end-users. So far, we provided one user story as example. You are required to add at least 4 more user stories.

- Managers
- Administrators
- Taxi Drivers

Managers (User Stories)

The team has already identified one user story for the managers user groups (shown below). You are required develop more user stories to build up your Product Backlog.

- **As** a manager, **I need** an intuitive application with a user interface **so that** I can visualize, analyze and compare taxi-driving data such as the telematics data on trip safety and make informed decisions.

Appendix 2: Managing Your Scrum Project

As the team have 4 months to carry out the project, we aim to conduct **Bi-weekly Scrum Report** (the team is going to record their progress, deliverables, and challenges).

The team also needs to decide on who will take on the various project roles, such as Data Translator, Data Engineer, Data Scientist, and Software Developer. (Your team may want to vote on this, take note that the roles can also be rotated if the team wishes so).

Data Translator: Aligning the team and the business side, analyzing the domain.

Data Engineer: Building data pipeline and data processing

Data Scientist: Machine learning models and optimization

Software Developer: Build Web Application and graphical user interfaces.

More roles if you can think about…