School of Mathematics and Statistics
Applied Data Science (MAST30034)
Project 1: Quantitative Analysis

Due date: 22nd of August 09:00 AM (AEST)
Project Weight: 30%

Project Overview

This project aims to make a quantitative analysis of the New York City Taxi and Limousine Service Trip Record Data. The dataset covers trips taken in various types of taxi and for-hire vehicle services in the New York City area. The data in parquet format is directly downloadable from here, with corresponding usage guide linked here. You will need to choose a **minimum of** 6 months if working with Spark or 3 months if working with pandas from 2016 or later (ensure your data includes Zones, not coordinates).

Students will be required to prepare a self-contained report which must be **at most 8 pages including figures, excluding references**.

Project Expectations

Please refer to the Canvas Subject Overview for expectations and further information.

We understand that the page limit is strict and quite short. This project aims to get students to be able to concisely summarise information professionally. This is because the results of Project 1 will be used to allocate which project student groups get for Project 2 (Industry Project).

Lastly, we know that the best way to learn new tools is to use and apply them in a project - this is "the project". Please try your best, the tutor team will be here to support you where possible.

Project Assumptions

- Students are free to choose any software, language, or package that is deemed useful to complete this project, although it is *strongly recommended* that Python and Apache Spark be used.
- A Latex report template will be provided and students are **not allowed to change the margins or font size**. Students who prepare their document templates will be required to add margin commands to adhere to the requirements. Otherwise, there will be penalties.

- Students must maintain **a GitHub repository** with an appropriate and documented `README.md` file. A template repository has been provided for your benefit under Canvas → Modules → Project 1 Links → Templates via GitHub Classrooms.

- Students have the freedom of choice to select their timeline to analyze, the *type* of Licensed Taxi you wish to focus on (i.e Yellow vs Green Taxi, Taxi vs For-Hire Vehicles), and the choice of attributes for their area of study. Once again, make sure the time frame chosen is 2016 onward.

- Students may use any external datasets which are deemed sufficiently relevant to support the analysis and attributes of the study.

- The timeline and dataset must be sufficiently "large" to support your research goal. Students may subsample the data when visualizing or fitting a model (please state this in the report or you will be penalised), but, **must use the full distribution** when analyzing the distribution, aggregating attributes, or performing outlier analysis.

Report Format

The report must be **at most 8 pages** (including figures, excluding references), covering at least, but not limited to, the following items:

- First and foremost, there should be **no code present in the report**. Please see the sample solutions for examples.

- Identify the taxi dataset, external datasets, attributes, timeline, target audience, and relevant research goal. Justification is required for each point.

- Outline the high-level methodology and preprocessing for visualization and statistical modelling for the research goal. We will be reading your code for the detailed preprocessing steps, so make sure it is well commented on and described.

- Preliminary data analysis with interpretation and discussion.

- A modelling section with at least two contrasting models and approaches with relevant evaluation metrics.

- Make practical and realistic recommendations based on the final results for the identified audience.

- Tables and figures should be referenced where appropriate. Here are some examples: *"From (Figure 3) we find ..."* or *"... the Gini Impurity Metric [3] suggests that ..."* or *"(Table 3) shows the ..."*.

- Ensure that figures are reasonably placed and readable, as ineligible figures or tables will be ignored.

- Less is more, choose the information you present carefully. Irrelevant information will make the report hard to follow and lead to significant reductions in marks.

- Finally, the report should be proofread several times before submission to minimise grammatical and spelling errors.

The Latex template is available via Overleaf or found under Canvas → Modules → Project 1 Links → Templates. You can download the source code and upload the `main.tex` to Overleaf or copy the project under Menu → Actions → Copy Project (located top left corner) on Overleaf. If you wish to use your Latex template, ensure your margins and document class adhere to our requirements by adding the following commands:

- `\documentclass[11pt]`{article}
- `\usepackage[top=0.9in, left=0.9in, bottom=0.9in, right=0.9in]`{geometry}

GitHub Requirement

The GitHub repository template is available via GitHub Classrooms or found under Canvas → Modules → Project 1 Links → Repo Template. You must use GitHub Classrooms and not your own personal repository.

All repositories will be cloned, executed (run), and used during marking, so please **ensure the code is reproducible and readable**. For example, if a student uses Python and uses external libraries, then a `requirements.txt` for a `pip` installation should be provided, such that anyone can run the command, install the packages, and run the code without errors. Repositories that fail to run will incur a penalty.

Assessment

This project is worth 30% of your final grade with the following requirements:

1. If no external dataset is used OR the student has chosen an insufficient dataset size, then the maximum number of marks is limited to 22.5/30 marks.
   - For example, if a student achieved 28/30 overall without meeting the requirements, their mark will be reduced to a maximum of 22.5/30.
   - If for some very unexpected reason you are unable to parse more than 6 months of data with Spark (or 3 months with pandas), you must let us know in advance via email with your reasoning.
   - We will provide a JupyterHub server to students with insufficient resources in a first-in, best-dressed manner.
2. If the chosen external datasets are relevant, justified, and used to complement the research goal, then full marks are awarded.
   - Some examples of suitable external datasets may be ongoing sports events, protests, weather forecasts (such as the impact of snow), vehicle crashes, etc.
   - There are several sources and some may require web scraping or direct contact with the owner of the dataset. It is up to students to choose and find one.

Strictly speaking, more marks will be available for students who perform additional analysis, with the highest marks available for students who perform *exceptional analysis* by drawing upon several external resources.

Hurdle Requirement

There is a hurdle requirement for you to submit a working GitHub repository and report. We have provided a template GitHub and Latex report for your benefit. Please ensure you **do not leave this until the last minute** to sort out as the submission deadline is strict.

Marking Scheme

This is an approximate marking scheme. Students who just "tick the boxes" may not always get full marks. We have provided this to be transparent on the marking process and expectations.

Introduction (3 marks)

- 2 marks for an appropriate choice of timeline, type of Licensed Taxi, attributes to study, dataset shape, and target audience, **with convincing justification for each of the choices made above**.
  - If you are using an external dataset, you must also clearly specify the details and provide a link to the dataset as a reference.
  - No marks will be awarded if there is a lack of justification.
  - All assumptions should also be introduced and stated in the Introduction.
- 1 mark for a single paragraph detailing the high-level overview of your methodology. Essentially, a "TLDR" of your contribution.

Preprocessing, Analysis, and Geospatial Visualisation (8 marks)

- 2.5 marks for suitably conveying the high-level preprocessing steps. If you would like to give it a small section, that is also fine. We recommend a simple dot point format (and remember that this is a summary, we will read through the code for an in-depth discussion).
- 6.5 marks in total for analysis of the attribute(s) related to your area of study. You should carefully consider the story you want to tell and the relevancy of the presented information. For instance, if your main message is that pickup locations play important roles in taxi drivers' overall revenue. Then presenting info about only tips will be a poor choice. Ensure you justify every step taken for full marks. Your presentation will need to include the following aspects:
  - 2.5 marks for initial outlier analysis, discussing the distribution, relevant imputations for `NULL` values, and summarised findings of interest for your chosen attribute(s).
  - 2 marks for describing some of the relationship(s) present between attribute(s) of "interest". You may consider interaction for this part.

- ○ 2 marks for discussion, particularly, if a certain visualization raises an "interesting" area for further analysis or results in the lack of anything "interesting" for further analysis.

Statistical Modelling (5 marks)

- 1 mark for clearly specifying and justifying at least two chosen models with your chosen attribute(s).
  - Make sure to list out your assumptions and ensure the attribute(s) are suitable.
  - Be very careful with the attributes i.e continuous vs ordinal vs categorical. Ensure your model is suitable.
  - If your model or algorithm was covered in a subject listed in the Subject Overview, you only need to reference and state it. Otherwise, please provide a brief introduction to your model with appropriate references.
  - Predictions should use future data (i.e using 2018 to predict 2019).
- 4 marks for analyzing their study goal by combining information from two models
  - 2 marks for analysis and/or comparison of models. This can be done through proper model refinement, feature selection, error analysis, model evaluation, or any suitable technique.
  - 2 marks for interpretation and discussion of the model for your study goal. If you conduct predictions, you should discuss their implications here.

Recommendations (4 marks)

- 4 marks for at least two sound recommendations for your target audience with the supporting evidence from previous sections.
  - For example, you should try and combine the findings of your analysis and model to give recommendations.
  - Recommendations should be practical. If it does not make sense to implement it, do not recommend it.

Report Writing and Code (10 marks)

- 2 marks for figures and tables with readable font and figure sizes.
- 3 marks for being able to convey the ideas and analysis (i.e through the use of correct and consistent grammar, spelling, citations, references, and report structure).
- 2 marks for code that runs without any issues and is maintained in a neat repository.
- 3 marks for readable code (i.e markdown cells, in-line comments, good variable names, reasonably adheres to PEP8 pylint or flake8). **This will be assessed especially for the preprocessing and modelling steps.**

Possible Additions (Maximum of +2 marks, capped at 30/30)

There will also be certain areas of possible additions for outstanding reports:

- Exceptionally well-written reports that make the reader go "wow".
- Recommendations and/or analyses that are realistic and feasible to deliver.
- Outstanding visualizations that are consistently great throughout the report.
- High code quality and readability that makes it easy to understand and run with no issues.

Possible Deductions (Maximum of -3 marks)

There will also be certain areas of possible deductions depending on the issues found:

- Insufficient quality of report writing (i.e the flow of the report was verbose or confusing, difficult-to-read sections, etc).
- Incorrect logic or breach of a business rule that has been stated in the Data Dictionaries.
- Overly verbose paragraphs in report and code.
- Insufficient commenting and detail in code for preprocessing and modelling sections.
- Figure sizes that are far too small.
- Significantly difficult-to-read code.

If you would like feedback on your code or report, please ask your tutor at the end of the tutorial during question time.

Submission Details

- Report submissions must be made via Turnitin on Canvas in PDF format written using Latex. We will not be accepting and marking any other format.
- Your final code must be in the GitHub repository and submitted on Canvas as a link.**Any submission without a GitHub link will fail this component**.
- Late submissions will incur a deduction of 10% (3 marks) per 24 hours past the submission deadline. If you submit late, you **must** email Calvin Huang (head tutor) at calvin.huang@unimelb.edu.au with your reason.

Extension Policy

If you have a valid reason with proof to request an extension, you **must** email Calvin Huang (head tutor) sufficiently before the submission deadline. Requests for extensions are not automated and will be carefully considered on a case-by-case basis. You **must** provide sufficient supporting evidence such as a medical certificate. Additionally, we will consider your `git` commits from your repository to illustrate the progress made on the project until the date of your request.

Academic Honesty

You are reminded that **all submitted project work and code** in this subject is to be **your work**. Automated similarity checking algorithms will be applied to compare submissions against all students, previous works, and known public sources. It is the University policy that cheating by students **in any form is not permitted** and that work submitted for assessment purposes must therefore be the independent work of the student concerned. Failure to comply may result in an Academic Honesty meeting with the faculty, with further escalation to the Academic Board depending on the severity.

To mitigate the risks of breaching Academic Integrity, please **cite and attribute all references and code functions** where applicable. For your report, you may choose any citation style listed on The University of Melbourne Recite page so long as you use it consistently.

Getting Started

(This is an example approach for the bare minimum marks.)

1. You could perform some basic geospatial visualizations on the Taxi data, compute descriptive statistics, and analyze summary statistics for your chosen attributes.
2. Then, you might formulate a relevant research goal and identify your client/stakeholder for your quantitative analysis.
3. Following this, you can build a Statistical Model to explain relationships between your input and response variables or use a Machine Learning model to classify/predict an attribute of choice.
4. Afterwards, you might investigate the correlation and feature relevance between your attributes, refine your model, and highlight key findings backed by your statistical analysis.
5. Finally, you should summarise and give recommendations to your identified clients or stakeholders.

In the event your results are unexpected or lead to unanticipated results, you should aim to discuss why they occurred and what it entails. This scenario happens quite commonly, so it's still in your best interest to make recommendations that support your unexpected results!

Additional Tips

If you're still unsure of how to start the project, try going through some of the materials and methods covered in the prerequisite subjects. Depending on the choice of Statistical Model or Machine Learning Algorithm, you may need to perform some creative feature engineering or transformation on the dataset.

For example, consider the scenario where your data is linearly separable through the use of a transformation or kernel trick:

- Consider performing some descriptive analysis before fitting your model to identify issues with your data such as linear separability, missing values, outliers, etc.

- For supervised learning models, consider the linear separability of your data. When there is linear separability, some models perform well (i.e SVM), whereas some models (i.e Logistic Regression) can fail to converge. The kernel trick may be used to induce linear separability.

- You should also correctly standardize/normalize your dataset depending on the model used.

- Penalised Regression Models such as Ridge (ℓ_2) and LASSO (ℓ_1) tend to perform poorly if the feature space is much smaller than the number of instances or if the attributes are not standardized.

- Consider performing feature engineering to generate more useful features. Do not perform it excessively though as it may lead to overfitting.

Final Tips

- Start this Project as soon as possible. It is up to you to spend as little or as much time as possible on this subject.

- You should aim to write your report professionally, assuming that an employer or client is paying you a salary or daily rate to conduct this analysis.

- Make sure you use a virtual environment or a new clean environment for development. Students are recommended to either use macOS or Linux for development. Windows users are recommended to use Windows Subsystem for Linux (WSL2).

- If you have too much data in a visualization, you can conduct sub-sampling to help increase the scope of data you can cover. Remember, you shouldn't have to describe your visualization in an overly verbose manner.

- Explain your handling of missing/unreasonable data and why any missing data does not undermine the validity of your analysis. You should report and justify the approximate size of data that has been removed.

- When you are trying to make comparisons between figures and tables, make sure your measurement is of the same scale (i.e do not compare miles to kilometres).

- Always tell the reader what to look for in tables and figures. Be as factual and concise as possible when reporting your findings with references where appropriate.

- If necessary, define unfamiliar concepts and provide the appropriate background information with references to aid your work.

- Good luck!