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

Due date: Friday 13th August 11:59am AEST
Project Weight: 35%

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 CSV format is directly downloadable from here, with corresponding usage guide linked here. You will need to choose a minimum of 3 months (≈ 4.5GB) for your project.

Please note that the datasets will have different coordinates structures depending on the year, so take this into account when choosing the timeline. For example, datasets in 2015 will include coordinates in latitude & longitude, whilst a dataset in 2018 will have aggregated zones. Students are also expected to find external datasets for their chosen coordinate representation if applicable.

Students will be required to prepare a self-contained report which must be **between 5 to 8 pages and between 1500 to 2000 words written in Latex** (excluding references). There will be penalties for exceeding the maximum word count, and you can check your word limit on overleaf.

Project Expectations

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

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 or R is used.
- A Latex report template will be provided and students are **not allowed to change the margins or font size**. Students who prepare their own document template will be required to add margin commands to adhere to the requirements.
- Students will also need to 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.

- Students have the freedom of choice to select their own timeline to analyze, the *type* of Licensed Taxi you wish to focus on (i.e Yellow vs Green Taxi's, Taxi vs For-Hire Vehicles), and the choice of attributes for their area of study.

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

- The timeline and dataset must be sufficiently "large" to support your research goal with a **bare minimum of three months without any sub-sampling** for preliminary analysis. Students may subsample the data when visualizing or fitting a model, but **must use the full set** when analyzing the distribution, aggregating attributes, or performing outlier analysis.

Report Format

The report must be **between of 5 to 8 pages and between 1500 to 2000 words written in Latex** (excluding references), covering at least, but not limited to, the following items:

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

- Outline the methodology and preprocessing for visualization and statistical modeling for the research goal. There should be **no code present in the report**.

- Analysis, interpretation, and discussion of findings.

- 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.

- 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 own 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 Hurdle Requirement

The GitHub repository template is available via GitHub Classrooms or found under Canvas → Modules → Project 1 Links → Template. If the template repository is not used, students must ensure their repository has a suitable `README.md` detailing the repository structure, scripting language(s) used, dependencies, and packages.

All repositories will be cloned and run, so please **ensure the code is reproducible**. 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 35% of your final grade with the following requirements:

1. If no external dataset is used OR the student has chosen less than three months for their taxi dataset, then the maximum number of marks is limited to 28/35 marks.
   - For example, if a student achieved 32/35 overall without meeting the requirements, their mark will be reduced to a maximum of 28/35.
   - If for some reason you are unable to parse more than three months of data, you must let us know in advance via email with your reasoning.
2. If the external datasets chosen 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 forecast (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 last minute** to sort out as the submission deadline is strict. There is a combined hurdle of 25/50 across your Project 1 and Assignment 1 results.

Marking Scheme

This is an approximate marking scheme. Students who just “tick the boxes” may not always get full marks.

Introduction and Data Selection (4 marks)

- 0.5 marks each (2 marks in total) for an appropriate choice of timeline, type of Licensed Taxi, attributes to study, and target audience.
  - If you are using an external dataset, you must also clearly specify the details and provide a link to the dataset as a reference.
- 1 mark for convincing justification for each of the choices made above.
- 0.5 marks for stating any assumption(s) made.
- 0.5 marks for stating the approximate dataset shape.

Preprocessing Steps (5 marks)

- 1 mark for correct use of imputations with justification.
- 2 marks for adequate investigation of outliers with justification.
- 2 marks for clearly stating the preprocessing steps in a concise manner.

Analysis and Geospatial Visualisation (10 marks)

- 2 marks for preliminary analysis on the attribute(s) related to your area of study. Note, preliminary analysis is typically at the single attribute level without considering interaction.
- 3 marks for describing some of the relationship(s) present between attribute(s) of "interest".
- 3 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.
- 2 marks for readable figures and tables, with suitable captions.

Statistical Modelling (6 marks)

- 1 mark for clearly specifying and justifying the choice of a single Statistical Model or Machine Learning Algorithm for either general interaction, prediction, or classification, with a suitable set of attributes.
  - If your model or algorithm was covered in a subject listed in the Subject Overview, you only need to reference it. Otherwise, please provide a brief introduction to your model with references.
  - Predictions should use future data (i.e using 2018 to predict 2019).
- 3 marks for the correct procedure for model refinement and/or feature selection with justification.
- 2 marks for correct use of evaluation metrics to conduct error analysis on the models' performance, with a rational investigation as to why the model performed better than, worse than, or as expected.

Recommendations and Discussion (4 marks)

- 4 marks for sound recommendations and discussion for the target audience with supporting evidence.

Report Writing and Code (6 marks)

- 4 marks for being able to clearly convey the ideas and analysis (i.e through the use of correct and consistent grammar, spelling, citations, references, and report structure).
- 1 mark for code that runs without errors (where applicable) in the submitted GitHub repository (hurdle).
- 1 mark for quality code that is readable (i.e commented, good variable names, and adheres to PEP8 or Google's R Style Guide).

Submission Details

- Report submissions must be made via Turnitin on Canvas in PDF format written using Latex.
- Your final code must be in the GitHub repository and submitted on Canvas as a link. **Any submission without a GitHub link will fail to meet the hurdle requirement**.
- Late submissions will incur a deduction of 10% (3.5 marks) per 24 hours past submission deadline. If you submit late, you **must** email the subject instructor Akira Wang at akira.wang@unimelb.edu.au with your reason.

Extension Policy

If you have a valid reason with proof to request an extension, you **must** email your Subject Instructor Akira Wang (akira.wang@unimelb.edu.au) 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 onto 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 own individual 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 it 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 kilometers).
- 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!