School of Mathematics and Statistics
Applied Data Science (MAST30034) 2024
An End-to-End Data Science Project

Due date: Monday 19th of August 11:59 pm
Project Weight: 30%

Author: Akira Takihara Wang, Liam Hodgkinson

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 `PySpark` or 3 months if working with `pandas` from 2016 or later (ensure your data includes Zones, not coordinates). `PySpark` is the expectation here; ensure you have obtained permission from your tutor to use `pandas`.

Students will be required to prepare a self-contained report which must be **6-8 pages including figures, excluding references, and written using LaTeX**. Please **do not** submit any other format written in Word or Google Docs, we are expecting a compiled `pdf` written in LaTeX. There are no exceptions.

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, consider this to be **"the project"**. Please try your best, the tutor team will be here to support you where possible. Sample solutions will be released at the start of Week 2.

Note: Students in prior years have often found themselves underestimating the time commitment required for this project. Be sure to start it ASAP; you should aim to have all of the results by the start of Week 4, so you can spend the remaining time writing.

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 `PySpark` 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. **We have been very clear here so do not submit any other document that was not written up in LaTeX.**

- 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 should 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" and justified to support your research goal. **Students may subsample the data when visualizing or fitting a model (you must state and justify 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 **between 6-8 pages** (including figures, excluding references, written in LaTeX), 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 datasets, external datasets, attributes, dataset shape, data timeline, target audience, and relevant research goal. **Clear and convincing 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 you include comments, docstrings, and markdown cells so it is readable even to a five year old child.

- Preliminary data analysis with interpretation and discussion.

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

- Make practical and realistic recommendations based on the final results for the identified audience. Recommendations must be supported by the analysis completed in the report and should not be generic. For example:

    - *"Taxi drivers can earn more money by taking more trips"* is not a good example.
    - *"Taxi drivers should aim to work their shifts at the airport during the day as we saw in our analysis above. This is due to XYZ and ABC reasons..."* is a good example.

- Tables and figures should be referenced where appropriate. Here are some examples:

    - *"From (Figure 3) we can see ..."*
    - *"... the Gini Impurity Metric [3] suggests that ..."*
    - *"(Table 3) shows the ..."*

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

    - If you look at the image in the report and it looks squashed, blurry, or hard to read, do not add it in. If you cannot read it, we cannot either.
    - If the font size is too small we will ignore it.

- 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. You have a plethora of online tools available, as long as you do not plagiarise we are okay.

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 referred to 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. If you are unsure about this, search it up and follow a guide or ask your tutor. Repositories that fail to run will incur a penalty.

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.

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 capped at 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 `PySpark` (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.

Marking Scheme

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

Introduction (4 marks)

- 1 mark for an *interesting* introduction to the problem. The first sentence is very important here, make sure that it is concise, precise, and engaging.
- Good discussion of data set timeline (0.5 marks); taxi type (0.5 marks); features of interest (0.5 marks); target audience (0.5 marks).
- **Must have convincing justification for each of the choices made above**.
  - As part of this assessment, it is your job to convince us that your choice is correct and valid, not the other way around.
  - No marks will be awarded if there is a lack of convincing justification. Please come up with a good reason even if you do not believe in it as it is critical that you can convince us that your choices are correct.
  - If you are using an external dataset, you must also clearly specify the details and provide a link to the dataset as a reference.
  - All assumptions about the dataset or business rules must be introduced and stated in the Introduction.
- 1 mark for detailing the high-level overview of your methodology. Essentially, a "TLDR" of your contribution. It is on you to be clear and specify all details, not on us to infer it and connect the dots. More detail is better than less here.

Preprocessing (3 marks)

- 1.5 marks each for the taxi dataset and another external dataset (may be more than one, but will be marked as one), suitably conveying the high-level preprocessing steps in dot points. Remember, keep it short and concise and only list out the important stuff.
  - If you are filtering, removing, or joining datasets, you *must specify the dataset shape at each step*. You will lose 0.5 mark if you do not do this. This will be verified with your code.
  - If you are removing records, you must justify and explain it or you will be penalised. For example, if you are removing all credit card payments because you are analysing tipping, then cite the business rule from the provided Data Dictionary and TLC Website.

Analysis and Geospatial Visualisation (5 marks)

- 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 analysis about only tips is irrelevant and a poor choice. Ensure you justify every step taken for full marks. Your report will need to include the following aspects:

- 1.5 marks for outlier analysis, discussion of the distribution, relevant imputations for `NULL` values, and summarised findings of interest for your chosen attribute(s).
- 1.5 marks for describing the important 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.
- Hint: For analysis and discussion, you should try following a "From (Figure X) we can see ABC, therefore DEF, perhaps because of XYZ" formula. We don't need students to explain the obvious, we are interested in reading your analysis and reasoning.

Statistical Modelling (5 marks)

- 1 mark for clearly specifying and justifying at least two chosen models with your chosen input and output attribute(s).
  - List out your assumptions and ensure the attribute(s) you have chosen are suitable.
  - Be very careful with the attributes i.e continuous vs ordinal vs categorical. Ensure your model is suitable. We will penalise incorrect usage of features.
  - 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) i.e do not use something like k-fold Cross Validation across 2019 and 2020 on a random shuffle.
- 4 marks for analyzing their area of research by combining the findings 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.
  - Hint: We are not interested in hearing about your model improving accuracy by a few percent. We want to hear how your model is going to support your target audience. For example, a taxi driver is not going to care about a Neural Network that predicts tipping. A taxi driver will however, be interested in a model that tells them where to go for more money.

Recommendations (4 marks)

- 4 marks for at least two sound recommendations for your target audience with the supporting evidence from previous sections.
  - 2 marks each: 1 mark for the recommendation, and 1 mark for a good justification / explanation.

- Be sure to combine the findings of your analysis and model to give recommendations.
- Recommendations should be practical and not generic. If it does not make sense in real life to implement the recommendation, do not recommend it.
- Generic recommendations that can be inferred using common sense will receive **no marks**. You have done hard work and analysis, make sure to use those findings to back your recommendations!

Report Writing and Code (9 marks)

- 2 marks for figures and tables with appropriate font size, figure sizes, choice of colour, etc.
- 5 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). Please proofread and use online tools to make sure your report is of high quality.
- 1 mark for quality code that runs without any issues and is maintained in a neat repository.
- 1 mark for excellent and readable code (i.e markdown cells, in-line comments, good variable names, docstrings for functions, 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 32/30)

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

- Exceptionally well-written reports that make the reader go *"damn that's actually good"*.
- Recommendations and/or analyses that are realistic and feasible to deliver and be considered for production.
- 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, capped at 0/30)

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 and TLC Website.
- 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 hyperlinked in the report. **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 the head tutor with your reason.

Extension Policy

If you have a valid reason with proof to request an extension, you **must** email the course coordinator 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

Okay, so this project spec is quite diabolical and intense. We understand you will be quite overwhelmed, and that's fine, the tutor team is here to make sure you get through it and learn everything you need before you graduate. It's part of the learning process so you are actually ready for your future career. So, we've written a short guide on how you might get started.

1. Start off by going through the Jupyter notebook for the first workshop and looking at the data. We recommend you also read the FAQs on the TLC Website and download the data dictionaries.

2. Then, carefully read through this spec and highlight or underline the important parts so you understand the requirements.

3. Then, figure out what kind of story you want to tell. For example, who is going to care about this analysis? Who do you want to help? Who's going to be paying you a salary to do this work? As a graduate of the Data field, it's really good to pretend to be your audience so you can analyse the data and help them.

4. For example, if your target audience is a Taxi Driver, then you should consider what life is like for them. What are my working hours? Where do I live? Which areas should I go for more potential rides? Do I want to avoid traffic?

5. Once you have figured it out, then you can start considering your timeline. For example, do I want to consider the impact of COVID? How about only post-COVID? How about only pre-COVID?

6. Then, you should consider okay, what are the attributes that related to my research area? What other information should I bring through? For example, should I look at sporting events, raves and music festivals, national public holidays, weather, and so on. These will be your external datasets.

7. Now, you can try aggregate your dataset timeline and present them on a geospatial visualization, compute descriptive statistics, and analyze summary statistics for your chosen attributes.

8. Then, dive deeper. Do some reaserch, talk to your fellow students, ask questions. Do some critical thinking and analysis. Always remember, is my target audience going to care about this analysis? If the answer is no, rethink your area of analysis.

9. Following this, you can build a Statistical Model to statistically explain relationships between your input and response variables or use a Machine Learning model to classify/predict an attribute of choice.

10. Afterwards, you might investigate the correlation and feature relevance between your attributes, refine your model, and highlight key findings backed by your statistical analysis.

11. 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 are here to learn and develop life-long skills.
- 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!