# DSA3361 Inferential Data Analytics

## Project Instructions

AY2022/23, Semester 2

## Introduction

The aim of the project is to practice some of the following skills:

- To integrate the course knowledge (and beyond) and apply it to real-life problems.
- To finish the pipeline of building up a machine learning model, starting from data cleaning/manipulation to model training and model evaluation.
- To simulate some real-life scenario in R and make data-informed decisions.
- To self-learn new topics (as needed) and have more practice in coding (in R).
- To communicate and collaborate with team members, and solve the problem(s) for potential users/stakeholders.
- To be able to break down a large problem into achievable steps.

The following is a suggested timeline:

| | |
|---|---|
| Week 6 – Recess Week | • Each team to decide on the general topic(s) (and the datasets if applicable).<br>• Each team to submit a 1-page proposal to the instructors. This proposal will be used as a discussion piece for feedback from the instructors, which will be useful to scope your project neatly for the final assessment.<br>• Your proposal submission is not graded.<br>***Deadline: 12pm, 29 Sep 2022 (Thursday)*** |
| Week 7 – Week 8 | • Each team to gather feedback from the instructors.<br>• Refine your project work and have another round of discussion with the instructors. (This can be an on-going activity till the end of the semester)<br>• Begin your main project work - set deadlines and deliverables for each team member. Do consider buffer time to consolidate your team's findings into a coherent project submission. |
| Reading Week | • Submit the required documents online. You can find more info under "Project submissions" (*See below)*<br>• Complete the peer-and-self-evaluation (*in Canvas*)<br>***Deadline: 12pm, 17 Nov 2022 (Thursday)*** |

## Project submissions (20% + 5% Bonus)

This course involves two branches: discrete-event simulation and regression analysis. For project, each group need to select one branch, find some relevant problem, and focus on it.

a) If you select "simulation", you need to decide one real life scenario, make necessary assumptions and simplifications, collect data to estimate the rates, build model(s), and utilise the model(s) to make data-informed decisions.

b) If you select "regression", you could pick some of the following tasks, which is included in the typical machine learning pipeline.
E.g., find/collect some datasets, make necessary data preparation (data cleaning: handle missing value, duplicates, incorrect values, outliers; feature engineering: make necessary transformations and/or craft new variables, etc.), conduct Exploratory Data Analysis (e.g., examine multicollinearity, correlation, visualization, summary statistics, etc.), build and train the model(s), evaluate the model's performance, practice model selection and feature selection, interpret the model(s) and make business implications.
Note that the data preparation and the exploratory data analysis (EDA) could be done iteratively. For example, you could first have some EDA analysis, followed by some data cleaning, and then generate more data visualisations (based on the cleaned dataset).

You can find more info under "Sample Projects". (*See Appendix)*

### Group submission

At the end of the project, each group will have to submit:

1. *An R Markdown file (.rmd)*, which be used to generate the full PDF report of the project, including code, charts, descriptions, and analysis. (Take reference to WS1.rmd)

2. *A 10-minute video (.mp4)*, uploaded to our class' Team group. The whole group can take turn to explain the PDF report that is generated by the rmd file.

### Individual submission

Everyone will have to submit:

1. *A 2-page Individual report*:

   One part is to provide a brief overview of the development of the project from your perspective. For example, how the problem was approached, how the work was assigned and how the difficulties were overcome as a group and as an individual. This part should also include a short review of the contributions of each group member.

   The other part is to serve as a technical manual of the work done. For instance, if you work on regression models, you can explain the models attempted, the data cleaning done, and how features/variables selection was conducted. If you work on simulation models, you can explain how the trajectory mimics the real-life scenarios, how the replications and optimisation were done, and what are the data-informed decisions. In particular, you should focus on the part that you contributed the most.

Note that the above submissions only focus on your chosen topic (e.g., simulation-related). To encourage everyone to practice more, we also recommend each group to practice coding on the other topic (e.g., regression-related), and you will be rewarded as well.

## Optional: Group submission (5% Bonus)

1. *An R Markdown file only. (No other document is required)*

As this is an optional task, you are not expected to spend time in data collection or building some new models. You are expected to utilise the course materials (including code), make some variations, and generate the new conclusions. For example,

a) If your primary submissions are regarding simulation, your group can simply choose one dataset in the appendix (more will be uploaded to MS Team later), propose some research or business question(s), mimic the code (in videos and in workshops), finish the key steps in the machine learning pipeline, make relevant discussions, and generate some data-informed decisions.

b) If your primary submissions are regarding regression, your group can simply choose any of the existing examples introduced in the course, vary the code/model, and make some follow-up discussions. You can either vary the scenario, or introduce more features into the existing model, or optimise the model with more combinations of key factors.

The intention of the optional task is to encourage everyone to practice coding for both simulation and regression. It also facilitates your preparation for the final exam. Each group member shall take lead in coding at least one component: EDA and data cleaning; Model training and evaluation (model and feature selection); building the trajectory and the model(s) in simulation; run replications, evaluate the monitor data and optimise the model's performance.

Please do not hesitate to arrange a meeting with us and we can discuss how to proceed.

## Peer-and-Self-Evaluation

To encourage a positive team dynamic and effective collaboration, an online peer-and-self-evaluation form in Canvas will be made available throughout the Reading Week.

The inputs from a group will potentially have a bearing on the project score of every individual in the group. You can choose not to respond to the evaluation. However, we will take that to mean you think everyone in your project group contributed equally.

## Project rubrics

See next page.

| Aspect | Description |
|---|---|
| **Data Preparation** | For regression:<br>• data cleaning: erroneous points, empty data, outliers<br>• feature engineering: transformation to existing variables or creation of new variables are applied to improve the analysis and provide new insights.<br>• EDA analysis: data visualisation is clean, clear, concise and captivating.<br>For simulation:<br>• Collect the real-life data to estimate the key parameters of the model(s), e.g., the arrival rates, the activity time rates, etc. |
| **Analytics** | • Demonstrate a sound understanding of the course-related content, and an appropriate application of the course-related techniques.<br>• The analysis shows a level of quality, integrity, and competency that make conclusions impactful, generating a high level of trust.<br>• For simulation, the quality of models, various performance measures, replications, optimisation and rank of solutions.<br>• For regression, feature selection and model selection, model evaluation/performance, model interpretability, and business insights.<br>• The quality of code shall demonstrate the proficiency of the groups' coding skills. |
| **Interpretations and data-informed decisions** | • Correctness of interpretation of data analysis and the resulting findings.<br>• Good understanding of strengths and weaknesses of the data and the analysis.<br>• Recommendations for improvement of the dataset or methods used leading to more accurate and potentially generalisable interpretations.<br>• An appropriate interpretation of statistics involved: distributions, confidence intervals, p-values, etc. |
| **Oral Presentation and Story-telling** | • The telling of a compelling, captivating, and coherent story.<br>• Confidence and clarity in speaking.<br>• Well prepared presentation and planning.<br>• Able to present in a logical and interesting sequence. |
| **WOW factors** | • Insights/Ideas are beyond what is taught/expected in the module.<br>• Good understanding and correct demonstration of programming skills that go beyond what is covered in the module.<br>• Correct identification and application of concepts taught beyond the module, to the topic. |

|  | Outstanding | Proficient | Adequate | *Poor* |
|---|---|---|---|---|
| Marks | 17-20 | 13-16 | 9-12 | 0-8 |
| General quality of all project submission components | All or almost all of the rubric descriptions are fulfilled and demonstrated in a convincing and clear manner. | Most of the rubric descriptions are fulfilled and demonstrated in a somewhat convincing and clear manner. | Some of the rubric descriptions are fulfilled and demonstrated in a somewhat convincing and clear manner. | Only a few of the rubric descriptions are fulfilled and demonstrated in an unconvincing and unclear manner. |

*For illustration purposes only.*

# Appendix A: Sample Projects (for Primary Choice)

For each sample topic below, please do not hesitate to arrange a meeting with us and we can discuss how to proceed.

## Simulation Project – NTUC Fairprice Store Model (Checkout/Queue System)

### Problem description (Scenario)

You are the chief manager of one NTUC Fairprice store, the nearest one to your house. You are in charge of re-evaluating the efficiency of the store's checkout process, in terms of the customers' wait time and the utilisation rate of service staff. The key factors of the model include the number of the cashier-checkout counters, the number of the self-checkout counters (kiosk machines), and the number of service staff (for checkout service only).

You can design your own collection of KPIs (performance measures). There is only one general design principle, which is to control the customers' wait time and the utilisation rate of service staff at a reasonable level (not too high for customers and not too low for staff).

Your job is to propose an optimised plan, which includes the number of checkout counters (two types) and the staff scheduling for the week.

### Data collection:

To simplify the task, you can randomly pick two days of any week: one weekday and one weekend. For each the two days, you could randomly pick two 10-minute interval: one in peak hour and the other in non-peak hour. For example, peak hour at weekdays may include 7-9am or 6-8pm and non-peak hour may include 2-4pm. To collect the customers' arrival data, one team member could stand at the entrance and counter the number of arrivals per minute, during the chosen periods (40 minutes in total). Therefore, your dataset should include 40 observations, each of which indicates the number of arrivals per minute. If you wish to gather more observations, you could either extend each interval to 20-minute, or count the number of arrivals per 0.5 minutes. Just like WS2, we could estimate the arrival rates at different timing (peak/non-peak of weekdays and weekends).

Similarly, you may also need to design your own way to estimate the (distributional) rates of activity time (both shopping time and checkout time).

More hints and tips would be provided to the groups who choose this topic.

## Simulation Project – KFC or McDonald’s Queueing System

The queueing system varies between KFC and McDonald, and among branches with different sizes. You can pick your favorite brand and store that you visited most often, and you are the chief brand consultant.

Just like the previous example, your job is to propose a new/improved design of the queueing and checkout process, in terms of the number of the staff-service counters, the number of the self-service kiosk machines, the number of collection counter(s), and the number of service staff (optional).

More hints and tips would be provided to the groups who choose this topic.

# Simulation projects’ list

- NTUC Fairprice Inventory Models, which is to simulate the maintenance of goods’ inventory and its replenishment (delivery and refilling).
- Pharmacy Model.
- Queueing System at Theme Parks (USS).
- Queueing System at Polyclinic.
- Any example/model from the book “Simulation Modelling and Arena” by Rossetti.
- Any example at https://r-simmer.org/articles/

## Regression Project – Red Wine Model

winequalityreds.csv

### Data Description

This datasets is related to red variants of the Portuguese "Vinho Verde" wine, from [Cortez et al., 2009]*. This dataset is also available from the UCI machine learning repository.

You can utilise the dataset for training either classification or regression models. The classes are ordered and not balanced (more normal wines than excellent or poor ones). There are 11 predictor variables, from fixed acidity to alcohol, and 1 outcome variable, quality (scores from 0 to 10).

As a wine expert, your job is to determine which physiochemical properties make a wine 'good'!

### Some possible tips and directions:

1) If you focus is classification, the outcome variable could be transformed into a factor

(categorical ordinal variable).

2) Perform Exploratory Data Analysis on your data set prior to data cleaning.
3) Produce at least 2 different (types of) graphical plots (bar plots, scatter plots, histograms, boxplots, etc) involving one or more variables, and highlight any interesting discoveries.
4) Data cleaning, variable transformation, and feature engineering.
5) Explore the multicollinearity, and the dependency of variables (e.g., correlation matrix).
6) Model selection: model performance and model interpretability.
7) Feature selection: Regularisation (Ridge/LASSO/Elastic Net), or some other method.
8) Cross-validation.
9) Statistical inferences.

More hints and tips would be provided to the groups who choose this topic.

## Regression Project Lists:

We will create a folder in both Canvas and Team to provide more datasets, that are suitable for training machine learning models. (Diamond, housing price, car price, etc)

You are also welcome to find your preferred datasets, from the industry sector that you wish to enter.

Note that it is not recommended to read the complete reports/analysis of the same dataset, from any online source. You are strongly encouraged to make your own attempts and practices.

# Appendix B: Sample Projects (for the Bonus Submission)

## Simulation Project – STEM Mixer Example

In the video's example, there are only three types of students: the wanderers who left without joining the booths; the wanderers who continued to join the two booths; the non-wanderers who directly went to the two booths. One way to improve this is to add more types of students: the wanderers who only joined JHBunt and left afterwards; the wanderers who only joined the Malwart and left; the non-wanderers who only joined one booth; the students who joined the two booths at a different order. You are recommended to read the chapters 3 and 4 of the book "Simulation Modelling and Arena" by Rossetti. Also, feel free to have your own assumptions and variations to make the model closer to the real context.

## Simulation Project – Bank Example

Pls refer to the tutorial blogs by Duncan Garmonsway [**], for options of improving the bank models. For example, we could simulate the VIP customers by giving them higher priority, with or without preemption; we could simulate the balking and the reneging customers, who either refuse to join a long queue, or quit the queue in the halfway.

Feel free to talk to us for more tips and help.

## Regression Project:

For any of the provided datasets (in Canvas and Team), you can follow the earlier instructions to attempt some of the hints/steps. Also, you don't need to finish each step in a thorough way. For example, you could just generate two or three charts for the data visualisation; you could just trained two or three models, evaluate them and make a comparison. In principle, the work on the bonus submission should not consume more than 20% of your time devoted to this project.

## Reference:

[*] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[**] The Bank Tutorial: Part I and II, by Duncan Garmonsway.

https://r-simmer.org/articles/simmer-04-bank-1.html

https://r-simmer.org/articles/simmer-04-bank-2.html