This document showcases my engagement in projects and tasks involving data science/analytics skillset. Due to non-disclosure agreements, most projects and tasks are unavailable for publishing. Please kindly refer to this document for my data science/analytics experience details.

Data management, manipulation, and interpretation/understanding

[Overview]

In my master's thesis, I needed to **preprocess the large and separate raw data (more than 6 million records in total)** and **track** and **record all the cleaning processes to enhance the reproducibility** of the project output.

[Approach/solution]

- I used Python to complete the process and **documented the cleaning process visually by flowcharts**, including the checks on **data format** and **duplication**, **deletion**, and **aggregation** of rows and columns.
- I also **made a data dictionary** of the final dataset stating each variable's description, data type, and range/options.
- Then, I assessed the data quality by key categories during **exploratory data analysis** and **visualization**.
- When any abnormalities were found, I **fixed them by iterating** the aforementioned preprocessing methods.

Statistical modeling (in R programming)

- **Generalized Linear Model (GLM)**:

  [Overview]

  I **analyzed the tendency of women's employment status** concerning their marital status, income of the male member in the household, presence of children, and region of residence. The dataset was based on 1977 surveys of Canadian couples and families. The model was GLM in the binomial family since the dataset had both continuous and categorical variables.

  [Approach]

  - **Exploratory data analysis** (EDA) by observing the correlations
    (1) between the outcome variable (i.e., women's employment status) and input variables and
    (2) among the input variables
  - **Model-fitting in various scenarios**:
    (1) using all input variables,
    (2) using fewer variables,
    (3) using all input variables one of which interacts with other input variables, and
    (4) using fewer variables and one of which interacts with remaining input variables
  - **Evaluation on models** based on the p-value of **ANOVA** test, **AIC** score, and **AUROC** score

  [Outcome]

  Based on the analysis, the implications were (1) the strong association of the presence of children and (2) the slight association of male members' income to women's employment status. I **achieved high distinction** for this task.

- **Segmented time series** and **ARIMA**:

  [Overview]

  I **analyzed the media attention impact on dispensing contraceptives** (combined/simple contraceptives) using **R**. The dataset consisted of monthly rates (per 1000 women of reproductive age) of PBS-subsidized dispensing of combined and simple contraceptives between January 2013 and December 2016. The media attention peaked in the last week of May 2015.

  [Approach]

  - **Exploratory data analysis** (EDA) by decomposing each time series data (i.e., the combined or simple) to observe the trend, seasonality, outliers, stationarity, and autocorrelation
  - **Log-transformation** of the data for eliminating autocorrelation and non-stationarity
  - **Model selection** for each data based on the EDA (e.g., stationarity + no-autocorrelation → segmented time series, no-stationarity + autocorrelation → ARIMA)
  - **Model fitting** for each time series by iteratively testing different parameters
  - **Evaluation of time series changes**: step (interruption) and slope after media attention (= intervention).

- **Quantifying the above changes in tables** and **visualized by the actual time series against the counterfactual** (simulative plot if no intervention was present)

[Outcome]
The media impact was agreeable on the combined contraceptives based on the changes (in %) with the monthly dispensing rate confidence intervals from the step change. I **achieved distinction** for this task.

Machine learning (in Python)

- **Supervised learning**:
  - **[Overview]**
    I **developed the algorithms to predict the risk of diabetic patients' readmission to a hospital** after discharge. The scenario was deploying the prediction algorithm for a hospital home-visit care unit, given that the operation cost is higher for readmitted patients. The dataset was a simulative electric health record data with binary labels of readmission (i.e., yes or no) and provided as "clean" data for this task. The algorithms used were **logistic regression** and **random forest**.

    [Approach]
    - ✧ **Train/test split of the dataset** with stratifying along the labels (target variable)
    - ✧ Developing the following **pipeline** for **logistic regression** algorithm due to its **sensitivity to the value scales**:
      (1) variable transformation
      (2) training/validation
      (3) hyperparameter tuning
    - ✧ **Fitting** (training/validation) with GridSearchCV on the Scikit-Learn library
    - ✧ **Model evaluation** by f1 scores (i.e., 2 x (precision x recall) / (precision + recall)),
    - ✧ Observation of feature variables by **SHAP** (for general feature importance) and **LIME** (for feature importance of a single sample prediction)

    [Outcome]
    I chose the random forest algorithm as the final model, given the higher scores in f1 (0.6706). Although the f1 score was not high, I concluded that the **random forest algorithm was deployable** given the **high precision score with test data** and **81% higher cost efficiency** with home visits **per patient**. I **achieved distinction** for the task.

  - **[Overview]**
    I **developed machine learning models** (final model as a **decision support system**) **for early-stage screening of Parkinson's disease patients**. The dataset consisted of 252 subjects (188 patients and 64 controls) with 3 records for each subject. The algorithms used were as follows.
    - ✧ **Logistic regression**
    - ✧ **Random forest**,
    - ✧ **Gradient boosting machine** (GBM)
    - ✧ **Artificial neural network** (ANN)
    - ✧ **Ensemble models** (voting/ensemble classifiers)

    [Approach]
    - ✧ **Exploratory data analysis/data preprocessing** (e.g., data type check, categorical variable check for encoding, balance between patient and control numbers, train/test data split)
    - ✧ **Developing models** per algorithm (with pipeline construction, when necessary, e.g., for logistic regression)
    - ✧ **Model evaluation** by **AUROC**, **recall**, and **f1** scores for prediction performance and over/underfitting: minimizing the false negative rate as the top priority

    [Outcome]
    Based on the model evaluation criteria above (priority in order), the **random forest model** was the best (AUROC: 0.8300, recall:0.97, and f1: 0.8706) and **likely viable only for preliminary screening** purposes. I **achieved distinction** for this task.

  - **[Overview]**
    My master's thesis was **developing a new (convolutional) neural network-based mortality forecasting model integrating cause-of-death** information using **Python**. The dataset was the U.S. mortality data from 1959 to 2019, initially with >6 million records.

[Approach]

- ✧ **Preprocessing/cleaning** the fragmented raw datasets (more than 6 million records in total) with **track and record** of the process to **enhance the reproducibility** of the project output
- ✧ Clarifying and excluding the **potential outliers** through **exploratory data analysis** to avoid the potential "noise" while training the model
- ✧ Assessing **the model performance** by a **comparison** table showing **the training time** and **mean square error** (MSE) values and graphs of forecast vs. actual of each model that **visualized the performance per age group**
- ✧ Clarifying the spot for future studies by **mentioning what needed to be added to my thesis's scope**

[Outcome]

The new model outperformed most of the compared models (e.g., Lee-Carter model) from the past studies by the smaller MSE value. I also presented the result to UNSW professors and research fellows and **achieved high distinction**. (Sample: https://github.com/MannyAdc/ForecastModel_LC_ML)

■ **Unsupervised/semi-supervised learning**

➢ **[Overview]**

I **developed a classification model for the images of blood cells** infected or uninfected by malaria using **Python**. The image data was provided as the compressed pixel data, and labels for the images were provided (both infected and uninfected images, 13,779 each).

[Approach]

- ✧ **Developing an autoencoder** (feed-forward neural network-based, **unsupervised**) to assess its power to distinguish the image by using a limited portion of the whole dataset (8,819 uninfected cell images) as the training data, **aiming to develop a model faster and less training data**
- ✧ Assessing the performance through several **data visualization**: direct comparison of the actual/reconstructed images and the t-SNE cluster plot

[Outcome]

I presented how **my model could distinguish the images**. I **achieved distinction** for the presentation.

➢ **[Overview]**

I **developed an autoencoder with LSTM for electroencephalogram classification** among alcoholic and control patients using Python. The dataset was from UCI Machine Learning Repository and consisted of 122 patients (120 trials for each patient and 255-step time series for each trial).

[Approach]

- ✧ Narrowing the data amount to 30 patients with 30 trials (20 patients with 20 trials for model training) due to computational capacity on my platform (Google Colab)
- ✧ Data loading and cleaning from .gz files (per trial) to pandas data frames (saved as CSV)
- ✧ Constructing and training the autoencoder with LSTM cells at each layer
- ✧ Assessing the autoencoder's prediction values from test datasets of alcoholic and control patients

[Outcome]

The **difference in mean square errors** was distinct between the alcoholic and the control. Hence, my **autoencoder was able to distinguish the data**. I **achieved high distinction** for this task.

■ **Reinforcement learning**:

[Overview]

I developed a reinforcement learning (**Batch-Constrained Q-learning (BCQL)**, hence model-free) **algorithm for hypotensive patient management in the ICU** using **Python**. The dataset consisted of vital signs, lab tests, and treatments measured over 48 hours in 3,910 patients with acute hypotension, and no new and additional real-time data was available.

[Approach]

- ➢ Defining the **reward** function which labels the mean arterial pressure (a vital sign) in the next time step,
- ➢ Labeling **state** (of each patient at each timepoint) by **k-means clustering** (i.e., **unsupervised learning**)
- ➢ Computing a tabular state-action (treatment) **value function** (i.e., RL policy)
- ➢ **Evaluating the RL policy performance** against the clinical policy (i.e., simply the observation dataset)

[Outcome]
The developed BCQL algorithm **outperformed the clinical policy** based on the expected value of reward. I **achieved high distinction** for this task.

SAS

Please see https://hds-hub.cbdrh.med.unsw.edu.au/posts/2023-01-13-sas-cortex/ for details. It refers to,

- winning **1st place in the SAS Cortex Analytics Simulation 5-Day Challenge** in April 2022,
- participating in the **internship program at SAS Institute Australia** as the reward for the above, and
- achieving **SAS Certified Associate: Programming Fundamentals Using SAS Viya** in June 2022.

SQL

[Overview]
Using the relational schema (57 data tables) about the information (e.g., people, program/course/class enrolment, facility, organization) from my university, I developed **PostgreSQL codes** to generate **views**, **tables**, and **functions** which take user inputs based on the task requirements.

[Outcome]
I **achieved high distinction** for this task.

BI dashboard (business setting)

[Overview]
At Sysmex, I **developed and delivered a BI dashboard using SAP Business Objects 4.0 Web Intelligence**. This responsibility started as a project of re-engineering the financial data analysis/reporting, which initially required an entirely manual process in Excel sheets before the data analysis with frequent errors (e.g., incorrect copy/paste, inconsistent version control).

[Approach/solution]
I have provided the BI dashboard and complimentary data extraction/checking tools to standardize and semi-automate the data preprocessing and **visualize the financial data (e.g., by time, subsidiary, and business field)**.

[Outcome]
Although the data granularity was often insufficient to create sophisticated analytical outputs given that the individual transaction level data was not publishable for multiple stakeholders, I still **delivered the frequently missed observation points (e.g., trend, irregularity, potentially incorrect data, false abnormality) through BI dashboard delivery**. In addition, I was **awarded by the Executive Vice President (head of division) for the 50%+ reduction of the existing analytics workload** of stakeholders in Japan

Microsoft Excel (business setting)

At Sysmex, I developed **MS Excel-based analysis tools** complimentary to the BI dashboard; for example,

- intragroup sales/cost forecast by items (from global headquarter to overseas affiliates),
- master data check file for annual maintenance, and
- revenue/profit breakdown simulation in various currency rates.

The purpose was to enhance the flexibility of analysis operation while optimally standardizing the tools for efficient and consistent usage. Some used functions are **vlookup**, **hlookup**, **index**, **indirect**, **subtotal**, **concatenate**, **pivot table**, and **pivot graph**.

Data entry, administration, and migration (business setting)

[Overview]
At Sysmex, I **engaged in data entry, administration, and migration** as part of financial analysis operation re-engineering through BI dashboard development and implementation.

[Issues/challenges]
The monthly financial reporting data (base data) did not have the granularity required for complete analysis. Initially, the data gathering and processing schemes for those “non-base” data were neither standardized nor automated (e.g., excel sheets in inconsistent form, calculation error during consolidation).

[Approach/solution]
I have developed the **scheme and tools that partially standardized and automated the data entry and administration processes** through **templates** and **rules** for routine data updates. Along with the implementation, I **cleansed** the past "non-base" data to align the templates and rules and then **migrated it into the data warehouse**, where the BI dashboard retrieved the data. Overall, I **engaged in the tasks for around three years** including the routine operation.

Translating the numbers to insights and actions (business setting)

I consulted a confectionery store (client) who initially feared their business termination due to their significant staff shortage. I analyzed numerically and visually their P/L and BS to quantify the possibility, measures, and potential risk of CF shortage. It required me to estimate

- the minimum viable revenue,
- the operation cost and cash flow,
- production capacity for sales at the store and to their distributors, and
- the condition of the client's staff members.

I then advised the client that they could maintain their business by

- reducing the business day/hours,
- temporarily discontinuing sales to their distributors,
- clarifying the time leeway before the cash flow shortage, and
- collaborating with another advisory team for job postings.

As a result, the client could maintain their businesses.