# Project 2 - Singapore Housing Data and Kaggle Challenge

Welcome to Project 2! It’s time to start modeling.

Shoutout to **Ben Lim**, Instructional Associate for DSIF-SG-4 for compiling this project!

**Primary Learning Objectives:** 1. Creating and iteratively refining a regression model 2. Using [Kaggle](https://www.kaggle.com/) to practice the modeling process 3. Providing business insights through reporting and presentation.

You are tasked with creating a regression model based on Singapore Housing Dataset. This model will predict the price of a house at sale.

This Dataset is an exceptionally detailed one with over 70 columns of different features relating to houses.

Secondly, we are hosting a competition on Kaggle to give you the opportunity to practice the following skills:

* Refining models over time
* Use of train-test split, cross-validation, and data with unknown values for the target to simulate the modeling process
* The use of Kaggle as a place to practice data science

As always, you will be submitting a technical report and a presentation. **You may find that the best model for Kaggle is not the best model to address your data science problem.**

## Set-up

Before you begin working on this project, please do the following:

1. Sign up for an account on [Kaggle](https://www.kaggle.com/)
2. **IMPORTANT**: Click this link ([Regression Challenge Sign Up](https://www.kaggle.com/t/dc36375f24b2489cb84f3f1d76e25b98)) to **join** the competition (otherwise you will not be able to make submissions!)
3. Review the material and download data files on the [DSI-SG-Project-2 Regression Challenge](https://www.kaggle.com/competitions/dsi-sg-project-2-regression-challenge-hdb-price/overview)
4. Review the [data description](https://www.kaggle.com/competitions/dsi-sg-project-2-regression-challenge-hdb-price/data).

## The Modeling Process

1. The train dataset has all of the columns that you will need to generate and refine your models. The test dataset has all of those columns except for the target that you are trying to predict in your Regression model.
2. Generate your regression model using the training data. We expect that within this process, you’ll be making use of:
   * train-test split
   * cross-validation / grid searching for hyperparameters
   * strong exploratory data analysis to question correlation and relationship across predictive variables
   * code that reproducibly and consistently applies feature transformation (such as the preprocessing library)
3. Predict the values for your target column in the test dataset and consider submitting your predictions to Kaggle to see how your model does against unknown data.
   * **Note**: Kaggle expects to see your submissions in a specific format. Check the challenge’s page to make sure you are formatting your CSVs correctly!
   * **You are limited to models you’ve learned in class so far**. In other words, you cannot use advanced supervised learning models like XGBoost, Neural Networks.
4. Evaluate your models!
   * consider your evaluation metrics
   * consider your baseline score
   * how can your model be used for inference?
   * why do you believe your model will generalize to new data?

## Submission

Materials must be submitted by the date specified by your Instructional Team through your GitHub account repo shared with the Instructional Team.

Your technical report will be hosted on Github Enterprise. Make sure it includes:

* A README.md (that isn’t this file)
* Jupyter notebook(s) with your analysis and models (renamed to describe your project)
* At least one successful prediction submission on [DSI-SG-Project-2 Regression Challenge](https://www.kaggle.com/competitions/dsi-sg-project-2-regression-challenge-hdb-price) – you should see your name in the “[Leaderboard](https://www.kaggle.com/competitions/dsi-sg-project-2-regression-challenge-hdb-price/leaderboard)” tab.
* Data files
* Presentation slides
* Any other necessary files (images, etc.)

## Presentation Structure

* **Presentation Time: 30 minutes**
* Use Google Slides or some other visual aid (Keynote, Powerpoint, etc).
* Consider the audience. Assume you are presenting to a non-technical audience (real estate agents, property owners, etc.).
* Start with the **data science problem**.
* Use visuals that are appropriately scaled and interpretable.
* Talk about your procedure/methodology (high level, **CODE IS ALWAYS INAPPROPRIATE FOR A NON-TECHNICAL AUDIENCE**).
* Talk about your primary findings.
* Make sure you provide **clear recommendations** that follow logically from your analyses and narrative and answer your data science problem.

Be sure to rehearse and time your presentation before class.

## Rubric

Teaching team will evaluate your project using the following criteria. You should make sure that you consider and/or follow most if not all of the considerations/recommendations outlined below **while** working through your project.

**Note:** Presentation will be done as a group while codes will be prepared and submitted by each student.

**Scores will be out of 27 points based on the 9 items in the rubric.** *3 points per section*

| Score | Interpretation |
| --- | --- |
| **0** | *Project fails to meet the minimum requirements for this item.* |
| **1** | *Project meets the minimum requirements for this item, but falls significantly short of portfolio-ready expectations.* |
| **2** | *Project exceeds the minimum requirements for this item, but falls short of portfolio-ready expectations.* |
| **3** | *Project meets or exceeds portfolio-ready expectations; demonstrates a thorough understanding of every outlined consideration.* |

### The Data Science Process

**Problem Statement** - Is it clear what the student plans to do? - What type of model will be developed? - How will success be evaluated? - Is the scope of the project appropriate? - Is it clear who cares about this or why this is important to investigate? - Does the student consider the audience and the primary and secondary stakeholders?

**Data Cleaning and EDA** - Are missing values imputed appropriately? - Are distributions examined and described? - Are outliers identified and addressed? - Are appropriate summary statistics provided? - Are steps taken during data cleaning and EDA framed appropriately? - Does the student address whether or not they are likely to be able to answer their problem statement with the provided data given what they’ve discovered during EDA?

**Preprocessing and Modeling** - Are categorical variables one-hot encoded? - Does the student investigate or manufacture features with linear relationships to the target? - Have the data been scaled appropriately? - Does the student properly split and/or sample the data for validation/training purposes? - Does the student utilize feature selection to remove noisy or multi-collinear features? - Does the student test and evaluate a variety of models to identify a production algorithm (**AT MINIMUM:** linear regression, lasso, and ridge)? - Does the student defend their choice of production model relevant to the data at hand and the problem? - Does the student explain how the model works and evaluate its performance successes/downfalls?

**Evaluation and Conceptual Understanding** - Does the student accurately identify and explain the baseline score? - Does the student select and use metrics relevant to the problem objective? - Is more than one metric utilized in order to better assess performance? - Does the student interpret the results of their model for purposes of inference? - Is domain knowledge demonstrated when interpreting results? - Does the student provide appropriate interpretation with regards to descriptive and inferential statistics?

**Conclusion and Recommendations** - Does the student provide appropriate context to connect individual steps back to the overall project? - Is it clear how the final recommendations were reached? - Are the conclusions/recommendations clearly stated? - Does the conclusion answer the original problem statement? - Does the student address how findings of this research can be applied for the benefit of stakeholders? - Are future steps to move the project forward identified?

### Organization and Professionalism

**Project Organization** - Are modules imported correctly (using appropriate aliases)? - Are data imported/saved using relative paths? - Does the README provide a good executive summary of the project? - Is markdown formatting used appropriately to structure notebooks? - Are there an appropriate amount of comments to support the code? - Are files & directories organized correctly? - Are there unnecessary files included? - Do files and directories have well-structured, appropriate, consistent names?

**Visualizations** - Are sufficient visualizations provided? - Do plots accurately demonstrate valid relationships? - Are plots labeled properly? - Are plots interpreted appropriately? - Are plots formatted and scaled appropriately for inclusion in a notebook-based technical report?

**Python Syntax and Control Flow** - Is care taken to write human readable code? - Is the code syntactically correct (no runtime errors)? - Does the code generate desired results (logically correct)? - Does the code follows general best practices and style guidelines? - Are Pandas functions used appropriately? - Are sklearn methods used appropriately?

**Presentation** - Is the problem statement clearly presented? - Does a strong narrative run through the presentation building toward a final conclusion? - Are the conclusions/recommendations clearly stated? - Is the level of technicality appropriate for the intended audience? - Is the student substantially over or under time? - Does the student appropriately pace their presentation? - Does the student deliver their message with clarity and volume? - Are appropriate visualizations generated for the intended audience? - Are visualizations necessary and useful for supporting conclusions/explaining findings?

In order to pass the project, students must earn a minimum score of 1 for each category. - Earning below a 1 in one or more of the above categories would result in a failing project. - While a minimum of 1 in each category is the required threshold for graduation, students should aim to earn at least an average of 1.5 across each category. An average score below 1.5, while it may be passing, means students may want to solicit specific feedback in order to significantly improve the project before showcasing it as part of a portfolio or the job search.

### REMEMBER:

This is a learning environment and you are encouraged to try new things, even if they don’t work out as well as you planned! While this rubric outlines what we look for in a *good* project, it is up to you to go above and beyond to create a *great* project. **Learn from your failures and you’ll be prepared to succeed in the workforce**.