Capstone: Group

This project is an opportunity for you to consolidate what you have learned during the course and apply it to a business problem that you work on with others.

You will have **three days** of class period to work on this project. Your group will turn in one version of all files that you use.

Scenario

You are a data analyst at **WOW!**, a leading real estate agency in Singapore. The HDB resale market has become increasingly competitive, with a rise in demand and volatile pricing trends. The agency wants to empower their agents with a data-driven tool to:

- Offer competitive pricing advice for clients looking to buy or sell HDB resale flats.
- Accurately estimate HDB resale flat values for different flat types and locations.

Your manager informs you that a team of data analysts is being formed **to analyse the HDB resale market in Singapore and build a predictive model in a short turnaround**.

Learning Objectives

- Leverage HDB resale price data to identify the top factors and how they influence HDB resale prices for **WOW!**: **What are the key factors that the agents of WOW! should focus on delivering accurate pricing advice?**
- Utilise the data analytics workflow within an agile development framework.
- Transform data into useful information to help support business decisions.
- Gain experience with project management and associated tools.

Deliverables

Submit a 20-minute client facing presentation containing:

- An overview of your analysis for the management team (15-minutes).
- Retrospective presentation (5 minutes) on Project Experience including:
    - What you have learned from this project.
    - Any additional files that you used as part of your analysis.
    - Screenshot (or share link) to your team's Trello Board.

Data

There are three files:

- **train.csv** -- this data contains all of the training data for your model. The target variable (SalePrice) is removed from the test set!
- **test.csv** -- this data contains the test data for your model. You will feed this data into your regression model to make predictions.
- **sample_sub_reg.csv** -- An example of a correctly formatted submission to evaluate the model accuracy on the test data. Refer to the following link for submission instructions - https://www.kaggle.com/competitions/dsi-sg-project-2-regression-challenge-hdb-price/data

Data Dictionary:

- `resale_price`: the property's sale price in Singapore dollars. This is the target variable that you're trying to predict for this challenge.
- `Tranc_YearMonth`: year and month of the resale transaction, e.g. 2015-02
- `town`: HDB township where the flat is located, e.g. BUKIT MERAH
- `flat_type`: type of the resale flat unit, e.g. 3 ROOM
- `block`: block number of the resale flat, e.g. 454
- `street_name`: street name where the resale flat resides, e.g. TAMPINES ST 42
- `storey_range`: floor level (range) of the resale flat unit, e.g. 07 TO 09
- `floor_area_sqm`: floor area of the resale flat unit in square metres
- `flat_model`: HDB model of the resale flat, e.g. Multi Generation
- `lease_commence_date`: commencement year of the flat unit's 99-year lease
- `Tranc_Year`: year of resale transaction
- `Tranc_Month`: month of resale transaction
- `mid_storey`: median value of storey_range
- `lower`: lower value of storey_range
- `upper`: upper value of storey_range
- `mid`: middle value of storey_range
- `full_flat_type`: combination of flat_type and flat_model
- `address`: combination of block and street_name
- `floor_area_sqft`: floor area of the resale flat unit in square feet
- `hdb_age`: number of years from lease_commence_date to present year
- `max_floor_lvl`: highest floor of the resale flat
- `year_completed`: year which construction was completed for resale flat

- `residential`: boolean value if resale flat has residential units in the same block
- `commercial`: boolean value if resale flat has commercial units in the same block
- `market_hawker`: boolean value if resale flat has a market or hawker centre in the same block
- `multistorey_carpark`: boolean value if resale flat has a multistorey carpark in the same block
- `precinct_pavilion`: boolean value if resale flat has a pavilion in the same block
- `total_dwelling_units`: total number of residential dwelling units in the resale flat
- `1room_sold`: number of 1-room residential units in the resale flat
- `2room_sold`: number of 2-room residential units in the resale flat
- `3room_sold`: number of 3-room residential units in the resale flat
- `4room_sold`: number of 4-room residential units in the resale flat
- `5room_sold`: number of 5-room residential units in the resale flat
- `exec_sold`: number of executive type residential units in the resale flat block
- `multigen_sold`: number of multi-generational type residential units in the resale flat block
- `studio_apartment_sold`: number of studio apartment type residential units in the resale flat block
- `1room_rental`: number of 1-room rental residential units in the resale flat block
- `2room_rental`: number of 2-room rental residential units in the resale flat block
- `3room_rental`: number of 3-room rental residential units in the resale flat block
- `other_room_rental`: number of "other" type rental residential units in the resale flat block
- `postal`: postal code of the resale flat block
- `Latitude`: Latitude based on postal code
- `Longitude`: Longitude based on postal code
- `planning_area`: Government planning area that the flat is located
- `Mall_Nearest_Distance`: distance (in metres) to the nearest mall
- `Mall_Within_500m`: number of malls within 500 metres
- `Mall_Within_1km`: number of malls within 1 kilometre
- `Mall_Within_2km`: number of malls within 2 kilometres
- `Hawker_Nearest_Distance`: distance (in metres) to the nearest hawker centre
- `Hawker_Within_500m`: number of hawker centres within 500 metres
- `Hawker_Within_1km`: number of hawker centres within 1 kilometre
- `Hawker_Within_2km`: number of hawker centres within 2 kilometres
- `hawker_food_stalls`: number of hawker food stalls in the nearest hawker centre
- `hawker_market_stalls`: number of hawker and market stalls in the nearest hawker centre
- `mrt_nearest_distance`: distance (in metres) to the nearest MRT station
- `mrt_name`: name of the nearest MRT station
- `bus_interchange`: boolean value if the nearest MRT station is also a bus interchange

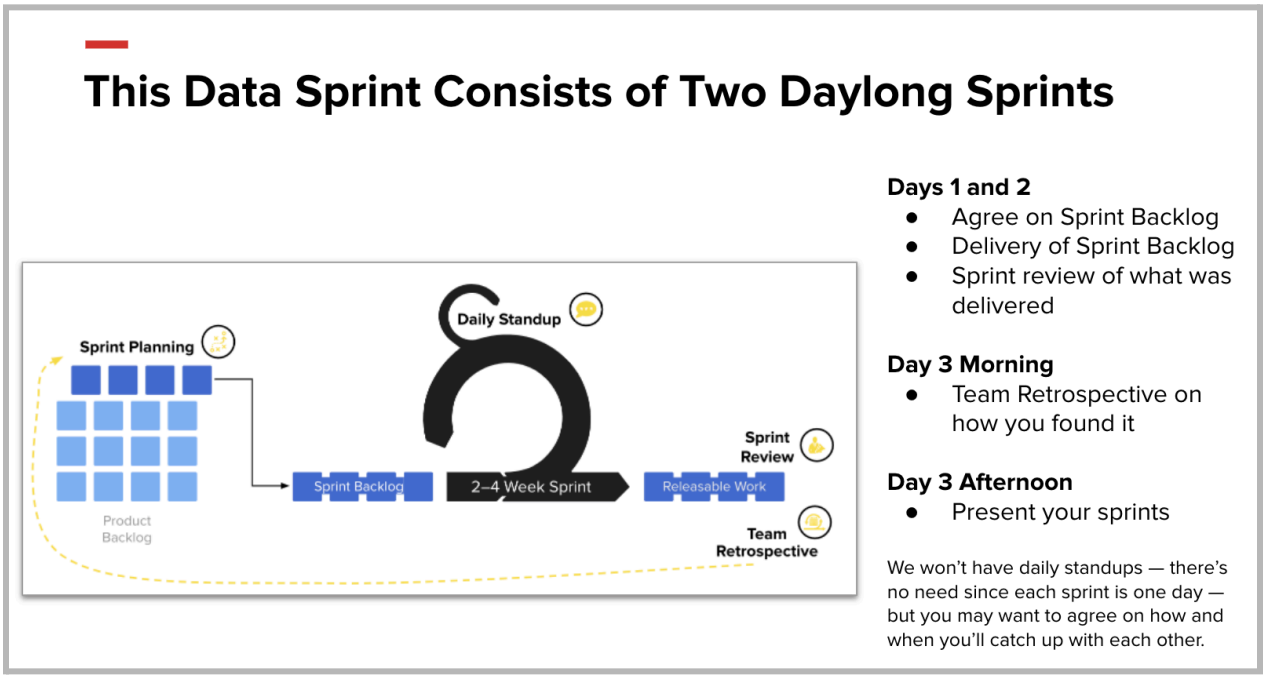
- `mrt_interchange`: boolean value if the nearest MRT station is a train interchange station
- `mrt_latitude`: latitude (in decimal degrees) of the the nearest MRT station
- `mrt_longitude`: longitude (in decimal degrees) of the nearest MRT station
- `bus_stop_nearest_distance`: distance (in metres) to the nearest bus stop
- `bus_stop_name`: name of the nearest bus stop
- `bus_stop_latitude`: latitude (in decimal degrees) of the the nearest bus stop
- `bus_stop_longitude`: longitude (in decimal degrees) of the nearest bus stop
- `pri_sch_nearest_distance`: distance (in metres) to the nearest primary school
- `pri_sch_name`: name of the nearest primary school
- `vacancy`: number of vacancies in the nearest primary school
- `pri_sch_affiliation`: boolean value if the nearest primary school has a secondary school affiliation
- `pri_sch_latitude`: latitude (in decimal degrees) of the the nearest primary school
- `pri_sch_longitude`: longitude (in decimal degrees) of the nearest primary school
- `sec_sch_nearest_dist`: distance (in metres) to the nearest secondary school
- `sec_sch_name`: name of the nearest secondary school
- `cutoff_point`: PSLE cutoff point of the nearest secondary school
- `affiliation`: boolean value if the nearest secondary school has an primary school affiliation
- `sec_sch_latitude`: latitude (in decimal degrees) of the the nearest secondary school
- `sec_sch_longitude`: longitude (in decimal degrees) of the nearest secondary school

Skills You'll Use

This project will let you try out the key skills we have learned throughout the course across the data analytics workflow, including: Framing questions with stakeholders, cleaning and processing data, visualising data, communicating your data, and modelling your data. And all this by working with others in an agile team.

It will require you to utilise a combination of Excel, SQL, Python, Power BI, and/or Tableau skills as you work through the data analytics workflow to analyse potential investments in the Singapore HDB resale market. However, in this sprint, you will also work within an agile framework.

Structure of the Sprint



- The data sprint will be a three-day 'sprint simulation': A hypothetical scenario which is based on a real-world HDB resale price market.
- You will be put into groups and will work in those breakout groups throughout the day.
- You will run a three-day sprint, and then wrap up the sprint at the end of day three and present your findings and what you have learned from the process in the morning of day four.
- Instructors will be playing the role of product owner and scrum master, and will help prioritise your work and coach you. The product owner has already created a product backlog and estimated the effort it will take for each item.
- You need to agree on the sprint backlog (high-level milestones) to complete with your group each day; you will also need to agree on a strategy for the day and a plan to divide up the work amongst you.
- You will conduct a sprint review with your product owner at the end of days one, two, and three to show them what you have managed to complete from the sprint backlog.
- We ask that you approach these three days with a creative analytical mindset!

Background

- A house represents the biggest investment for most households. Whether you're a home owner or a tenant, your mortgage payments or monthly rent will likely consume a significant portion of your income. Meanwhile, with some of the highest property prices in the world, real estate as an asset class poses an out-sized impact on the macroeconomic environment in Singapore.
- Therefore, being able to accurately value current housing prices will not only facilitate market transactions by providing valuable guidance for all market participants (be it home owners, home buyers, landlords, tenants or banks that underwrite mortgages), but also provide useful insights for policy makers and government authorities in understanding the current state of the economy.
- This is, however, an incredibly difficult endeavour. Real estate is notoriously known for its illiquid nature, owing to the fact that every single home is unique; there's literally only that ONE home at that exact location!
- You have been tasked with analysing the HDB resale price market and outlining recommendations for building an accurate HDB resale price predictor .

Daily Breakdown of Tasks and Deliverables

Preliminary Work

Project Introduction (90 minutes)

- Go through what the sprint involves, and the scenario.
- Discuss how you will divide responsibilities for delivering as much value from the backlog as you can in two days.
- Clarify any questions you have on how the sprint will work.

Preparation Session (90 minutes)

- Research the HDB resale market and housing trends in Singapore to understand the market dynamics, pricing factors, and potential investment opportunities. You will need this to effectively analyse the data. For your research, explore relevant government websites, real estate market reports, and academic articles.
- Investigate the provided data sets and review the data dictionaries.
- Review the scrum and agile workflow, and how you will use them in this data sprint.
- Review the product backlog, clarify anything you need with the product owner (Instructor), and be prepared to agree upon the sprint backlog for day one within the first hour of day one.

Day 1

Agree on the sprint backlog for sprint one (1 hour)

- Work out what you can accomplish as a team today, and any help you may need
- Ask the product owner for any questions to clarify items in the product backlog
- See the **tips for creating your sprint backlog** section below for help with structuring
- Agree on the sprint backlog with the product owner for sprint one (day one)

Team Work Tasks (rest of the day except for the sprint review)

- Complete your sprint backlog utilising Excel, SQL, Python, Tableau and/or Power BI to engage in the DA workflow.

Sprint Review (30 minutes)

- Complete a 10-minute show-and-tell of the work that was completed during the day with the product owner.
    - Summarise the team's findings and insights and any tasks that still need to be completed. These can be notes, and can be in any format you choose. Review visualisations and dashboards that were created.
    - Discuss any challenges or struggles that you experienced and how you overcame them.
    - Discuss further steps that need to be taken and any part of the sprint backlog that was not completed.
- Spend no more than 20 minutes preparing for the show-and-tell
- Submit your show-and-tell
- You do not need to submit any other files at the end of today

Day 2

Retrospective (30 minutes)

- In your team, reflect on day one and answer five questions:
    a. What went well?
    b. What didn't go so well?
    c. What have you learned?
    d. What still puzzles you?
    e. Do you need to change your ways of working at all for day two?
- Discuss and agree on the structure of day two, taking into consideration your reflections.

Agree on the sprint backlog for sprint two (30 minutes)

- Work out what you can accomplish as a team today, and any help you may need
- Ask the product owner for any questions to clarify items in the product backlog
- Agree on the sprint backlog with the product owner for sprint two (day two)

Team Work Tasks (rest of the day except for the sprint Review)

- Complete your sprint Backlog utilising Excel, SQL, Python, Tableau and/or PowerBI to engage in the DA workflow.

Sprint Review (30 minutes)

- Complete a 10-minute show-and-tell of the work that was completed during the day with the product owner.
    - Summarise the team's findings and insights and any tasks that still need to be completed. These can be notes, and can be in any format you choose. Review visualisations and dashboards that were created.
    - Discuss any challenges or struggles that you experienced and how you overcame them.
    - Discuss further steps that need to be taken and any part of the sprint backlog that was not completed.
- Spend no more than 20 minutes preparing for the show-and-tell.
- You do not need to submit any other files at the end of today

Day 3

Retrospective (30 minutes)

- In your team, reflect on day two and answer five questions:
    a. What went well?
    b. What didn't go so well?
    c. What have you learned?
    d. What still puzzles you?
    e. Do you need to change your ways of working at all for day three?
- Discuss and agree on areas of improvement that you can use for working together on day three.

Team Work Tasks (rest of the morning)

- Prepare a presentation that includes your team's insights and recommendations actions for the WOW! management team:
    - Review the team's findings, insights and recommended actions.
    - Discuss any challenges or struggles that you experienced and how you overcame them.
    - Discuss further steps that would be taken in the next sprint.
- Tip: You can pick a slide template from https://slidesgo.com/ and download it.

Presentations (afternoon)

- Each team will deliver a 25-minute presentation:
    a. Proposal for the WOW! management team (15 minutes)
    b. Lessons learned as a team from doing the project (5 minutes)
    c. Q & A (5 minutes)
- The intended audience of the presentation is technical and non-technical.
- Each team member will contribute to the delivery of the presentation.
- The pdf of the story will be submitted at noon on day three. All other files will be submitted by midnight on day three. Include a document that summarises all submitted materials.

Tips For Creating Your Sprint Backlog

Spend time looking at all the items in the product backlog. Their priority has been set by the product owner, but you can discuss this with them if you disagree with any of them.

The estimated relative priority of each item (Story Points) has been set by the product owner, who talked to various individuals in your department as they defined the backlog.

You need to take the product backlog, clarify anything you need to (which may involve the product owner defining some items in more detail, or even splitting them out into multiple items), and then agree with the product owner which ones you'll do in your first sprint.

For each item in the product backlog, consider:

- Is the item well written? Is it clear what's required?
- What would you need to do in order to complete the task?
    - What data would you need?
    - What data wrangling would you need to do?
    - What tools would you use? Excel, Python, SQL, Tableau, Power BI, etc.?
    - What are some possible issues you might face?
- Could you divide and conquer, or would you all need to work on it together?
- How long do you think it would take? Is the relative effort listed in the product backlog right?
- Do you think the item's priority is right?

Tips For Inspecting Your Data Sets

Start with ***loading the data*** into Excel or Python, or examine it using SQL, and get familiar with what you have (and what you don't have).

When inspecting the data sets:

- What does each column mean?
- Which features (columns) do you think will be the most useful?
- Which features (columns) do you think you won't need?
- Is there missing data? If so, how could you handle it?
- What other data would you like to get, to help your analysis? Is it easy to find and get? If so, is it complete?
- Conduct descriptive statistical analyses in Excel or Python.

Work out what data cleansing / wrangling is required

- Which data source(s) will you use to answer each question in your sprint backlog?
- What cleaning do you need to do to the data? For example, renaming columns, joining data, and handling nulls.

Create data table relationship(s) in order to run queries in SQL.
Conduct visualisations and dashboards in Excel, Tableau, Power BI or Python.

Explain your data wrangling process (if any) when you present to the investment committee. Note any important new columns you created in your data tool and any steps you took to clean the data, change data types, handle missing data, merge data, etc. Include comments throughout your work.

Good luck and have fun!