Overview of Data Files

This section documents the sources of our datasets.

Data Features	Description	Source
Resale Flats Transactions	Database which provides an extensive record of all new sale and resale property transactions from 1995 onwards and is made accessible through NUS Libraries	Real Estate Information System (REALIS) from the Urban Redevelopment Authority (URA)
Property Price Index (PPI)	The PPI provides an indicator of the market-wide property price levels and it tracks the movement of this price level across time	Real Estate Information System (REALIS) from the Urban Redevelopment Authority (URA)
Train Stations	Dataset representing MRT and LRT as geospatial point data in shapefile format	https://www.mytransport.sg/content/dam/datamall/datasets/Geospatial/TrainStation.zip List of Singapore MRT stations by planning area
Schools	Dataset containing general information of schools in Singapore such as postal code	School Directory and Information
Singapore Geographic Information	An extensive API consisting of geospatial as well as corresponding demographic data	OneMap Singapore
Crime Rate	Dataset documenting preventable crimes reported to at the different neighbourhood police centre	Preventable Crime Cases Recorded by NPC-Data.gov.sg List of Neighbourhood Police Centres | AFD

Modelling Data Files

This section provides a description of the respective datasets we used for our modelling processes.

	File Name	Description
Crime Cases	police centre_dataset.xlsx	List of neighbourhood police centres and their respective locations
	crime cases_dataset.csv	Records of reported crime cases in respective neighbourhood police centres from 2011 to 2018

Train Stations	mrt.sg.csv	Each row documents an individual MRT or LRT station. Interchanges have multiple rows, each representing the individual train lines that an interchange carries. Details include station name, station number, location, and the associated color of the train line (eg. Yellow for Circle Line).
	mrt_openingDates.xlsx	Records the opening dates of each MRT station.
	lrt_openingDates.xlsx	Records the opening dates of each LRT station.
	train_stations_data.csv	Preprocessed and compiled dataset to include train station information from the other files above. This is the output dataset from ***01A Train Data Compilation*** notebook under the Feature Engineering folder.
Primary Schools	general-information-of-schools.csv	Information like postal code, street addresses of schools from Primary to Tertiary level. We only extracted data of Primary schools.
	priSch_openingDates.csv	Records the opening dates and closing dates (if available) of primary schools. This dataset was filled manually as there was no available dataset online on opening and closing dates.
Property Price Index	property_price_index.csv	Records the property price index for each quarter from 2010 to 2020.
Property Specifications	Completion_Date_Manual_Fill.csv	Includes the year of completion for properties with missing completion date.
Feature Engineering and Final Preprocessing	preliminary_dataset.csv	Consists of information such as project specifications and the engineered features. All features are compiled into a single dataset for further preprocessing It excludes duplicated resale transactions and those that are out of the scope of our project.

		This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
	final_dataset.csv	Subset of the preliminary_data.csv with a smaller number of features. This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
	modelling_dataset.csv	Consists of all the variables that will be used for the model experimentation process after final preprocessing. This is the output dataset from ***01D Final Preprocessing*** notebook under the Feature Engineering folder.

Interface-Related Data Files

This section provides a description of the respective datasets we used for the building of our interface.

	File Name	**Description**
Crime Cases	average_cases_by_npc.csv	Get the average number of crime cases for each police centre. Dataset is used to display the average yearly number of crime cases for the nearest police centre to the user's inputted listing in Interface. This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
	police_centre_gdf.csv	Get the geo dataframe for police centres. Dataset is used to get the nearest police centre to the user's inputted listing in Interface.

		This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
Train Stations	train_gdf.csv	Get the geo dataframe for train stations. Dataset is used to get the nearest station, stations within 1km and train lines within 1km to the user's inputted listing in Interface. This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
Primary Schools	primary_sch_gdf.csv	Get the geo dataframe for primary schools. Dataset is used to get the nearest school to the user's inputted listing in Interface. This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
Geographic Data for Properties in preliminary_data.csv	historical_postal_code_area.csv	Get the geo dataframe for postal codes that are in our preliminary_data.csv. Dataset is used to get the planning area/latitude/longitude of the user's inputted listing if that listing is in our cleaned dataset before, so that we do not have to run OneMap API. This is the output dataset from ***01C Interface Datasets*** notebook under the Feature Engineering folder.
Centroid for Planning Area	area_centroid.csv	Get the centroids for each planning area. Dataset is used to get the planning area of the user's inputted listing.

		This is the output dataset from ***01C Interface Datasets*** notebook under the Feature Engineering folder
Past Resale Transactions	preliminary_dataset.csv	Consists of information such as project specifications and the engineered features. All features are compiled into a single dataset for further preprocessing It excludes duplicated resale transactions and those that are out of the scope of our project. This is the output dataset from ***01B Feature Compilation & Engineering*** notebook under the Feature Engineering folder.
	modelling_dataset.csv	Consists of all the variables that will be used for the model experimentation process after final preprocessing. Dataset is used to ensure that the columns of train dataset used during modelling and columns of user's inputted listing are the same to allow for prediction of price on the Interface. This is the output dataset from ***01D Final Preprocessing*** notebook under the Feature Engineering folder.