AY 2022/2023, SEMESTER 2

Group Project Final Report

Github Repository: https://github.com/Zifengzf/IS3107
Google Username: is3107v1@gmail.com
Google Password: is3107password

End-to-End Pipeline of Public Housing Data in Singapore

Project Group 3

Name	Matric Number
Kwong Wai Kit	A0214352R
Eugene Lim	A0217424L
Su Zifeng	A0222369A
Lee Chen Xi	A0220687B
Melvin Dio Haryadi	A0219799A

Table of Contents

Introduction

The Housing Development Board (HDB) is the sole provider of public housing in Singapore. These public houses offer affordable housing options for Singaporeans, with almost 80% of the population living in them. Given that housing is one of the largest purchases and investments that a household makes, it is crucial to make an informed decision about price trends across various locations and housing types. Our project aims to benefit prospective buyers and property agents that require a detailed analysis of housing price trends for specific categories of houses to make better data-driven decisions.

Dataset & Sources

In this project, we made use of several data sources to construct a combined dataset for our analysis. These include HDB resale and BTO statistics released by the Singapore government, financial statistics obtained from Yahoo Finance and Reddit comments in the Singapore subreddit.

HDB Resale Prices

Our primary dataset is HDB resale flat prices obtained from data.gov. It is updated daily and contains 897,607 records covering transactions from 1 January 1990 to date, as of 20-4-2023. Notably, only the month and year are specified for each transaction. Variables in tabular form are presented below.

Variable	Type	Description
month	String	Date of data record as "YYYY-MM"
town	String	Neighbourhood of flat sold
flat_type	String	Variable that categorises flats typically by the number of rooms in the flat, or other special categories e.g., EXECUTIVE, MULTI-GENERATION
block	String	Block number of flat
street_name	String	Street name of flat
storey_range	String	Range of storeys that the flat falls under, with 25 unique categories
floor_area_sqm	Integer	Floor area of transacted flat in square meters
flat_model	String	A different categorisation of flats which are less self-descriptive eg. MODEL A, NEW GENERATION
lease_commence_date	String	Date of start of 99-year lease as "YYYY"
resale_price	Float	Sale price of flat

Fig 1. Resale Transactions Dataset Description

Yahoo Finance Data

To enrich the resale transactions dataset with economic indicators, we incorporated financial data obtained from the Yahoo Finance API. There are 400 records in the dataset, dated from January 1990 to the present, as of 20-4-2023. Only the last datapoint of each month are retained to align with the resale transaction dataset. The table is as shown below.

Variable	Type	Description
Date	String	Date of data record as "YYYY-MM"
_10Y_Treasury	Float	10-year United States treasury yield in percentage
S_P	Float	S&P 500 market index tracking the stock performance of 500 of the largest companies listed on stock exchanges in the United States

STI	Float	Straits Times Index tracking the performance of the top 30 companies that are listed on the Singapore Exchange

Fig 2. Yahoo Finance Dataset Description

HDB BTO Data

To compare the trade-offs between purchasing a resale HDB outright and balloting for one via BTO, we included the HDB Price Range dataset from data.gov. This dataset currently contains 265 records and covers price ranges of Build-to-order (BTO) flats offered by HDB from April 2008 to March 2022, updated annually. The variables provided includes maximum and minimum selling prices across room types, towns, and years. The complete variable table is attached in the appendix under Figure 3.

Reddit Data

To supplement our statistical data with sentiments from potential buyers, we incorporated online comments from buyers regarding HDBs and BTOs on the r/Singapore Reddit page via the Reddit API. The 100 newest posts are extracted and each comment is uniquely identified by a comment id, alongside comment text, and date. The complete variable table is attached in the appendix.

Points of Interest

The main variables of interest to us are resale_price, town, flat_type, storey_range, floor_area_sqm, flat_model, remaining_lease from HDB Resale Prices & _10Y_Treasury, S_P and STI from Yahoo finance dataset. We wish to investigate the relationships between resale_price and the rest of the variables which define the flat coupled with the addition of financial variables to represent the economic state that flats were transacted under.

Our supplementary variables of interest are town, room_type, min_selling_price & max_selling_price from BTO Prices, and polarity_score from the reddit data. Even though it may not directly affect the price of resale flat, BTO Prices variables serve as good reference points to access the relative value of resale flats. Polarity score from reddit data also serves as a reference for the market sentiment at different points in time.

Motivation

During our exploratory data analysis, we observed large disparities in housing prices across towns after controlling for other factors. For Q4 2022, we observe the highest median prices for 4-room flats in Queenstown ($870,000) and the lowest prices in Jurong East ($465,000). Overall, we observe that location is a highly influential factor in housing prices.

Apart from location, we believe that a host of other factors are likely to be influential to housing prices. This includes factors such as year, flat model (5-room, Executive), floor area and the broader economic environment. We would like to verify the strength of the relationship between these factors and resale housing prices, both across time and different cross sections of the housing market.

For young couples purchasing their new home, a common dilemma is balloting for a BTO or purchasing a resale flat directly. We would like to evaluate the extent of the resale premium based on housing attributes, which would contextualise the trade-offs made in terms of cost and time.

Data Pre-processing

Resale Transaction Data

For resale transaction data, there are 2 main transformations to the provided dataset. First, the month column of format 'YYYY-MM' in the original dataset is separated into month and year. Second, we created an additional feature for remaining lease in years, derived from the lease commencement date and transaction year. These features were transformed to streamline the pipeline of data into the machine learning model, and dashboard for visualisation.

BTO Data

Financial Year	Town	Room Type	Min. Selling Price (S$)	Max Selling Price (S$)	Min Selling Price Less AHG/SHG (S$)	Max Selling Price Less AHG/SHG (S$)
2020	Choa Chu Kang	4-room	253,000	326,000	193,000	266,000
2020	Choa Chu Kang	5-room	0	0	0	0

Fig 3. Data Cleaning for BTO Dataset

The BTO dataset contained some rows of data where all the prices are $0, denoting that the combination was not offered. These rows were removed during pre-processing.

Financial Data

Yahoo Finance API returns the Open, High, Low, Close and Adjusted Close Prices, as well as Volume. We decided to retain only the Adjusted Close Price for 10Y Treasury Rates (^TNX), Straits Times (^GSPC) and S&P500 Index (^STI) as Adjusted Close Price accounts for corporate actions, such as stock split, mergers, and dividends. For historical data, we would retain the last data point of every month to be aligned with the labelling of the Resale Transaction Data. For the current month, we would use the latest available data point which would be updated every day.

Database Schema

Database Design

After the pre-processing stage, the datasets were uploaded to Google BigQuery and stored as tables. Due to the non-relational nature of BigQuery, primary and foreign keys were not required during schema specification. To enrich the resale transactions dataset for machine learning, resale transactions was joined with financial data on the fields of month and year, which is illustrated in the table below.

In BigQuery, we chose not to partition the data due to the low volume of data in our tables with a maximum of only 120 megabytes. Clustering was also not utilized as BigQuery only supports clustering on partitioned tables.

Database Preview

The diagram below shows the 4 tables and the relationships between them. The month and year attributes have a one-to-many relationship, where a single record in financial data corresponds to multiple transactions occurring in the same month and year. Snapshots of each table in BigQuery are provided in the appendix.

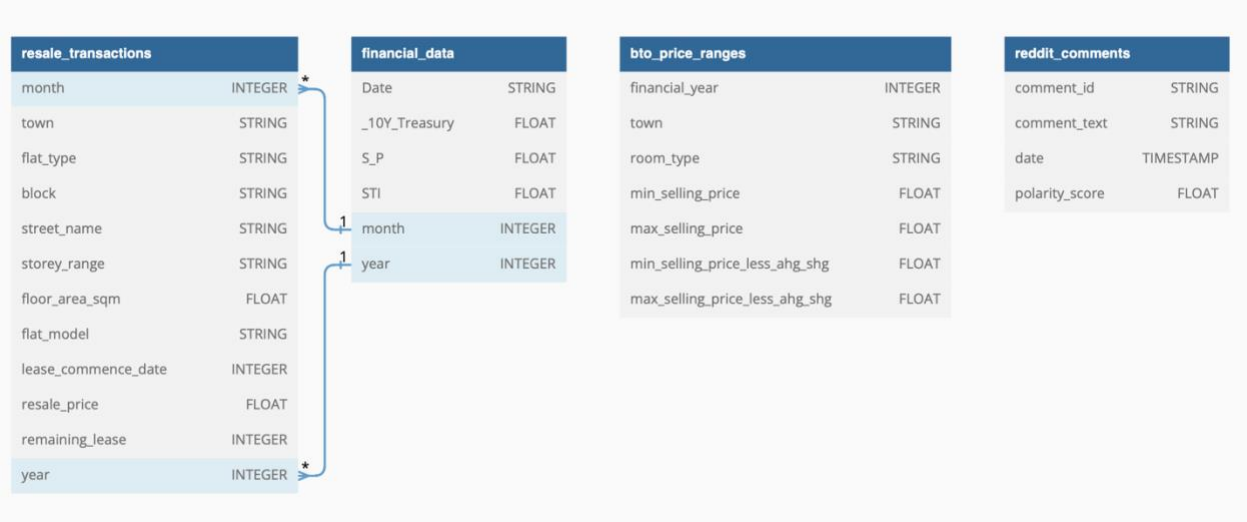


Fig 4. BigQuery Tables Overview

Data Pipeline

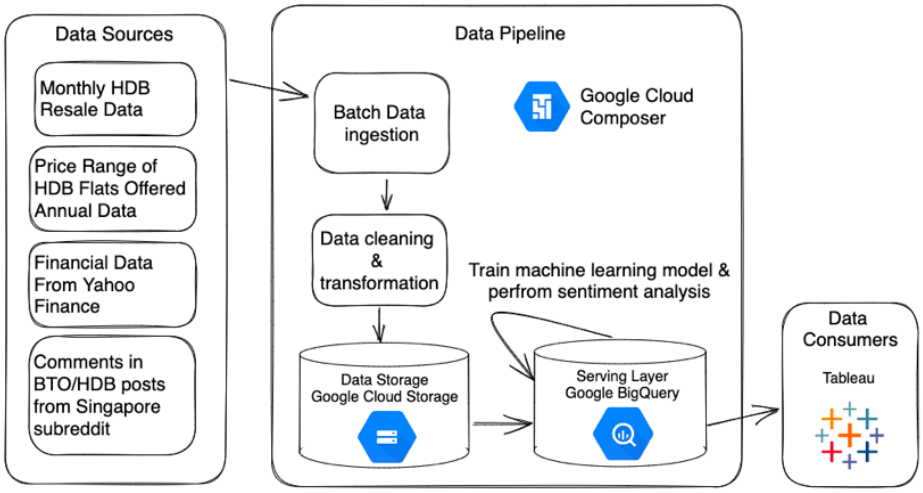


Fig 5. Data Pipeline Diagram

Platform Selection

Our data pipeline is primarily built on the Google Cloud Platform (GCP), using services such as BigQuery to serve queries and Cloud Composer for orchestration. Scalability, integration, and cost were the 3 main factors we considered for our choice.

GCP offers a highly scalable infrastructure well suited to handling growing data volumes and processing, a suite of integrated services that streamlines our development process, allow remote collaboration on the same instance, and offers free credits that offset development costs.

Whilst GCP provides an in-house business intelligence and data analytics platform, Looker Studio, we utilized Tableau for analysis and data visualisation instead. This is due to Tableau's richer feature set and compatibility with BigQuery through a simple connection and authorisation.

Service Selection

The services on GCP need to fulfil the functions of data storage and querying, machine learning and data pipeline orchestration. For data storage and querying, we relied mainly on Google Cloud Storage in conjunction with BigQuery, which provides all rounded functionality in storage, querying and analysis of data, given that most our data is structured. For machine learning, we opted for BigQuery ML. Given its integration with BigQuery, it allows for a simpler data pipeline by handling of data in place. With respect to orchestration, we opted for Cloud Composer, due to its ease of use as a managed Airflow service, and integration with other GCP services such as BigQuery.

Workflows

There are a total of 3 separate DAGs running different workflows on Google Cloud Composer. The first workflow relates to resale and financial data. The second workflow updates the BTO data in a similar fashion while the final workflow recreates the Reddit comments table with the newest comments.

Resale and Financial Data DAG

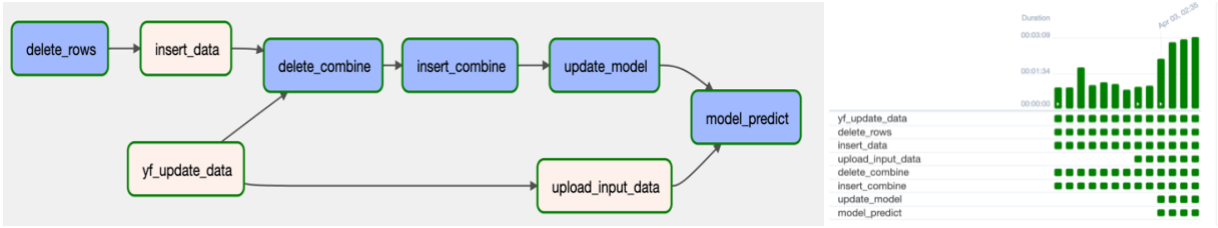


Fig 6. DAG of Resale & Financial Data Workflow

This DAG is responsible for updating the resale transaction and financial data tables, joining them to update the combined data table, training the machine learning model on the new data and making predictions for future resale prices. Execution for this DAG happens daily.

The rationale for performing deletion and insertion, instead of a simple append only method is due to the constraints of the API. The resale data API maintains an internal ordering of transactions that is based on year and town name, rather than strictly based on transaction date. Without a primary key to uniquely identify each transaction, we could only query the API for all transactions in the latest year and replace the existing records in the resale transactions table to ensure that there is no duplication. For financial data, we seek to align it with the resale transactions, retaining the last adjusted close price of the month, which must be updated as the DAG is run for each day. Once the individual tables are updated, we join them on the fields of month and year, before employing the same approach of deletion and insertion on the combined table. This approach is favoured over creating a new table from scratch due to computation costs.

With respect to the machine learning tasks, we generate a table of features to perform prediction on via upload_input_data, retrain our model on new data via update_model and perform prediction for resale prices via model_predict. Our main rationale for keeping machine learning and table updating in the same DAG is the dependency of model input features on financial data, model training on combined data and the relatively low training time of less than 2 minutes for the model. The final result of this DAG includes updated tables of resale transactions, financial data, combined data, model prediction inputs and model prediction outputs, which would be used for visualisation in Tableau.

BTO Data DAG

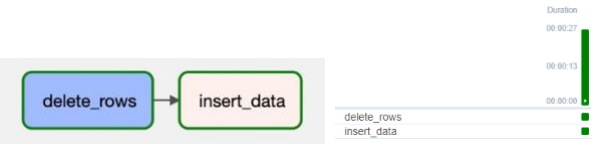


Fig 7. DAG of BTO Data Workflow

This DAG is responsible for updating the BTO data table, with a similar logic to the resale transactions table. Execution happens monthly due to the frequency of updates.

Reddit Comments DAG

Fig 8. DAG of Reddit Sentiment Workflow

The final DAG is responsible for updating the reddit comments table with the respectively comments and sentiment score. It first authenticates and obtains a token from the Reddit API, then performs a search query to get a list of the newest 100 posts. For each post, we query the comments using the post ID, perform sentiment analysis using TextBlob on the comments to obtain an aggregated polarity score. We then write the comment_id, comment_text, date, and polarity_score to BigQuery.

To prevent duplicates, we have decided to delete the table for reddit comments before insertion. Given the small size of the dataset, it is computationally faster to replace the entire table in a single operation than performing multiple filtering, deletion, and insertion. Having a record of historical comment data is also unnecessary as they would not be used for any further analysis.

Due to the low frequency of new comments, reddit data is processed as batch data executed daily instead of streaming data.

Run Times

The run times of each task of the three DAGs are attached in the appendix.

Sentiment Analysis

Objective

Our purpose for employing sentiment analysis is to provide summary of the general sentiment among the public regarding the housing market.

Sentiment Analysis

We have selected a pretrained model from the TextBlob library to extract the polarity scores for each comment. TextBlob has high predictive power as it is trained on a large corpus of textual data. In addition, TextBlob has a simple to use and intuitive API, which can easily integrate into our workflow. TextBlob also eliminates the need for pre-processing the raw textual data as it performs tokenization automatically, enabling a simpler and more efficient workflow.

Machine Learning

Objective

Our purpose for employing machine learning is to predict future resale housing price trends across towns, flat types, floor levels and economic environments proxied by interest rates and stock indexes.

Model Creation

BigQuery ML supports several machine learning models for regression, including linear regression, random forest, boosted trees, deep neural network, and Auto ML. Using SQL queries, we create models for all the options provided, except for Auto ML due to the extreme computation time (>2 hours). These models are then stored natively in BigQuery, which can be called by queries through the Google Cloud UI for testing or triggered as jobs in our airflow pipeline.

As part of the model creation query, we specify the ML model and feature list. Optional parameters for data split, feature processing and hyperparameter tuning, were omitted in favour of the default options provided by BigQuery ML.

Feature Selection

We selected US 10Y Treasury Rates, S&P500 and SPX Index, town, flat type, storey range and year as the features to predict resale price. These were the most significant features based on empirical testing.

Data Split

We performed a random split of 10,000 records for evaluation and the rest for training, as we have more than 50,000 records in our dataset. While we considered a commonly used 80-20 randomized split, we ultimately decided to limit the number of records for evaluation to minimise overfitting.

Feature Processing

For feature processing, we perform standardization for numerical features and one hot encoding for categorical features for all models except Boosted Tree and Random Forest. We find these transformations adequate for model training.

Hyperparameter Tuning

For the hyperparameter tuning algorithm, we relied on Google's Vertex AI Vizier, which uses a combination of advanced search algorithms such as Bayesian Optimisation with Gaussian Process. While options including random search and grid search were available, the Vizier algorithm performed better with substantially reduced model training times.

Model Evaluation

We evaluated the models along three dimensions, accuracy, speed, and computational cost. Our models were trained on data up to 2022 and evaluated using data from the past 6 months as the validation set.

Model Type	**Linear**	**Boosted Trees**	**Random Forest**	**DNN**	**DNN + Linear**
Training Time	2 min	10 min	10 min	10 min	10 min
R^2	0.733	0.782	0.791	0.475	0.463
Explained Variance	0.726	0.829	0.818	0.371	0.380
Processed Bytes	67.01MB	32.82GB	32.82GB	16.59GB	16.64GB

Fig 9. ML Model Evaluation

In terms of accuracy, boosted trees and random forest perform the best, followed by linear regression. However, due to the simplicity of the linear model, the training time and computation costs are far superior to any of the other models. In addition, linear regression provides a clear and interpretable relationship between housing features such as floor area against resale price, which can be easily evaluated when performing sanity checks of the model. We believe that the slight reduction in accuracy is a worthwhile trade-off for speed, cost, and interpretability.

Model Prediction

As part of the data pipeline, once the financial data table is updated, we generate a table of input features, which contains all permutations across the feature list, which would be used for predictions later. This includes financial data projections across different economic regimes, using the latest financial data as the baseline and projecting up to 5 years into the future.

Once the combined table of financial and resale data has been updated, we train a new linear regression model which replaces the old model and predicts the resale prices in the future based on the input feature table created. The predicted results are then saved to the output table in BigQuery, which is drawn on for tableau visualisation.

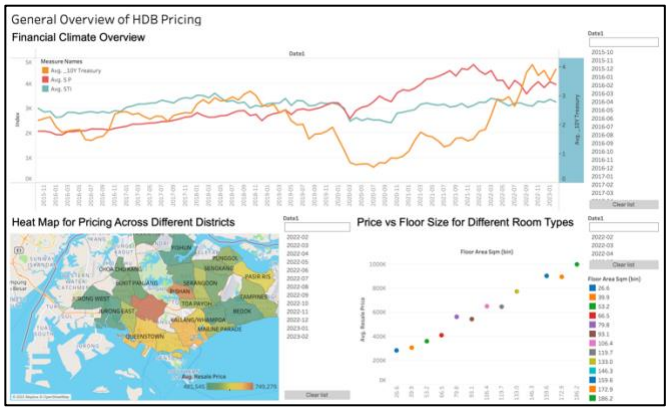
Data Visualisation

Recognising the difficulties that prospective homeowners face when making a purchase decision, we created a dashboard with 6 segments that provide both broad-based and precise insights about housing price trends. With the shift towards digitalisation since the pandemic, we also see potential for the platform to support property agents in engaging their clients with data driven decisions on their real estate purchases.

General Overview of HDB Pricing

Our landing page consists of 3 charts to summarise recent trends and developments in real estate industry and the general financial climate.

At the top of the dashboard, we have a line graph of the US 10Y Treasury Rates, S&P500 and SPX Index. Given that Singapore's S$NEER policy band largely follows that of the US policy rate decisions, we included US 10Y Treasury rates to provide a sensing of recent financing costs and the interest rate environment. Equity indexes, S&P500 and SPX Index are also included to provide colour on the general economic sentiment which is an important factor for users to determine their ability to finance their purchase. From the chart, we can see that we are facing record high interest rates while the equity indexes seem to trend lower indicating some bearish sentiments in the market. As such, it may be worthwhile to hold off their property purchases until future rate cuts.

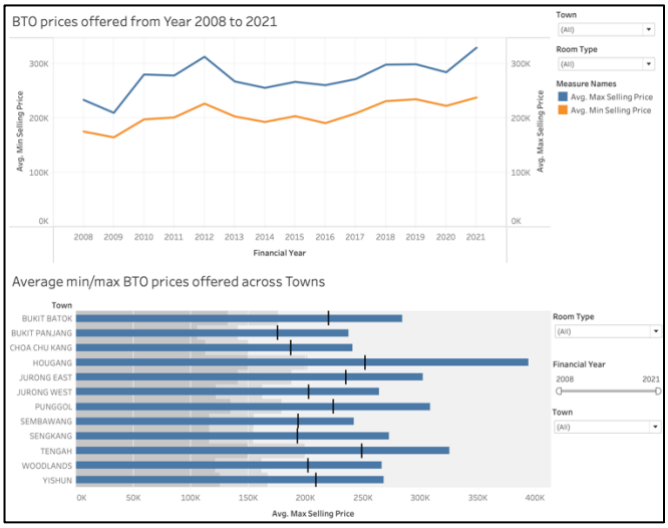


On the bottom left of the dashboard, we plotted a heatmap of historical resale prices by colour coding the different districts on Singapore's map to better facilitate users to determine a suitable property location within their budget. In general, property in central districts, such as Bukit Timah, are associated with higher prices compared to property in the outskirts. This is likely due to the high desirability of its proximity to the CBD and some of the most prestigious schools in Singapore. An anomaly would, however, be Pasir Ris with an average resale price of SGD 625k as compared to property in other outskirts of Singapore ranging around SGD 500k.

The third graph on the bottom right of the dashboard depicts a scatter plot of the average resale prices against the different floor sizes. This enables users to evaluate the relationship between resale prices and floor sizes. Generally, property with larger floor area is deemed as more desirable and commands a higher price. A noteworthy observation is that the increase in resale prices is more significant between 70 sqm and 80 sqm compared to other increments of 10 sqm.

Historical BTO Prices

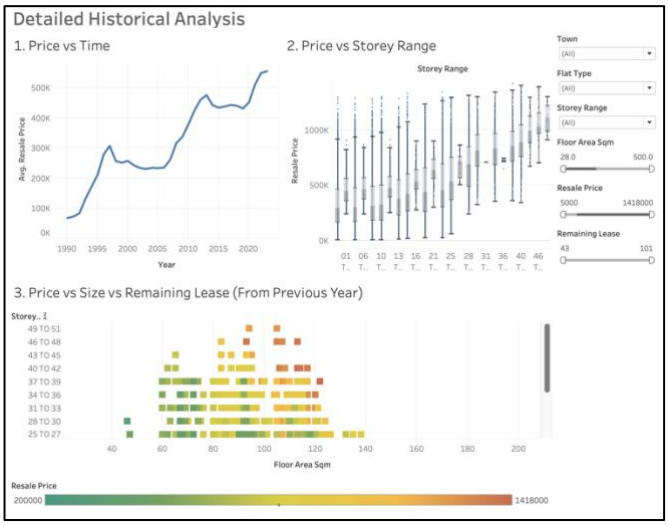
On this page, we included a line graph for the minimum and maximum selling prices of historical BTO. We can see a general uptrend in BTO prices, which makes sense given that Singapore is very land scarce and the increased demand for housing as population grows. However, it is interesting to see a sharp increase in the maximum selling price of BTO from 2020 to 2021 which would likely be because of the pent-up demand from the pandemic.



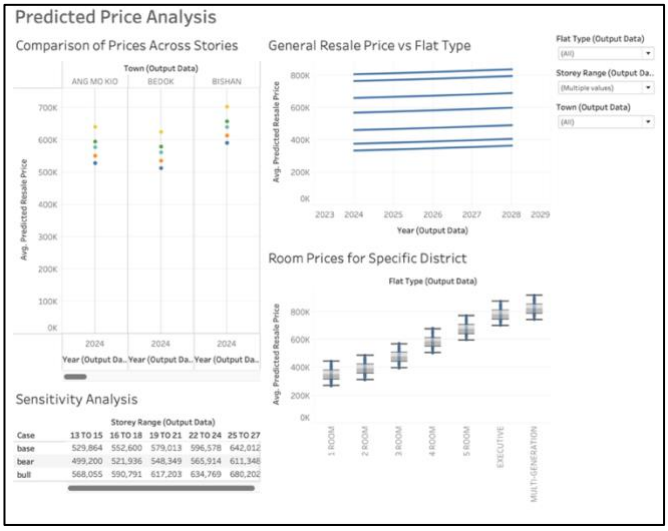
On the bottom of the page, we also included a horizontal bar graph to showcase the maximum and minimum price variations of the across the different towns. The maximum price is indicated by the bar while the minimum price is indicated by the black line.

Detailed Historical Analysis for Resale Data

For this dashboard, we feature a line graph, boxplot, and a coloured scatter plot to enable users to analyse the historical resale data better. After getting a better idea of the kind of property users would wish to purchase from the prior dashboards, they would be able to filter for specific towns, flat types, storey ranges, floor area and resale prices for a more thorough analysis.



Predicted Price Analysis



For the final dashboard, we present our findings of the ML model which is the predicted prices of resale property for the next 5 years. Our team included a dot plot graph of the resale prices of each flat type across the different selected town, a multi-series line graph of resale price data for each flat type, a sensitivity table for the price prediction with a bear, base and bull economic case, and a boxplot to showcase the variance of predicted prices across the different districts. All forecasted prices, except for those in the sensitivity table, would be based on the base economic case. An interesting observation would be that a HDB in Bedok within the 25 to 27 storey range is equivalent or slightly more expensive than that of a HDB in Bishan within the 16 to 18 storey range. As such, users could have a sensing of how much in terms of stories they would need to compromise should they prioritise a more central location.

Sentiment Analysis

We have decided to incorporate a simple sentiment analysis dashboard as one of our downstream applications. This dashboard can track the recent sentiments regarding HDB and BTO by checking comments from posts about the HDBs and BTOs in Reddit. The comments feed shows the recent comments along with hues that represent its polarity based on a particular date interval we can specify in the filter. We can also look at the average polarity score from the comments within the specified period. Moving along the dashboard, we can also see boxplots for the comments for specified periods. And finally, users can see the relationship between text length and polarity from our density plot.

Conclusion

Through our data pipeline, we have created a predictive machine learning model for resale housing prices across housing segments and economic environments. These predictions, alongside historical data extracted from various sources and sentiment analysis scores derived from the Reddit API, are used to create dashboards that showcases housing price trends and implications of the various features of a property on resale prices. Our end-to-end pipeline enables users to better able to understand the local property market and make a more informed choice on the specific property they wish to acquire.

Appendix

Dataset Descriptions

Data	Type	Description
financial_year	Integer	Financial year that the flat was offered in, given in "YYYY"
town	String	Neighbourhood of flat offered
room_type	String	Variable that categorises flats typically by the number of rooms in the flat
min_selling_price	Float	Minimum selling price of flat offered
max_selling_price	Float	Maximum selling price of flat offered
min_selling_price_less_ahg_shg	Float	Minimum selling price of flat offered after accounting for the maximum possible amount of Additional Housing Grant (AHG) or Special CPF Housing Grant (SHG), which are subsidies for eligible first-timer families, that a buyer can receive
max_selling_price_less_ahg_shg	Float	Maximum selling price of flat offered after accounting for the maximum possible amount of Additional Housing Grant (AHG) or Special CPF Housing Grant (SHG), which are subsidies for eligible first-timer families, that a buyer can receive

Fig 10. Price Range of HDB Flats Offered Dataset Description

Variable	Type	Description
comment_id	String	Unique identifier for comment
comment_text	String	Body of the comments
date	Timestamp	Date of the comment creation, in timestamp format
polarity_score	Float	Polarity score of the comment, calculated using TextBlob's polarity scoring algorithm

Fig 11. Reddit Comment Dataset Description

Database Previews

price_ranges QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW LINEAGE

Row	financial_year	town	room_type	min_selling_pric	max_selling_pric	min_selling_pric	max_selling_pric
1	2008	PUNGGOL	2 ROOM	82000.0	107000.0	0.0	0.0
2	2008	PUNGGOL	3 ROOM	135000.0	211000.0	0.0	0.0
3	2008	PUNGGOL	4 ROOM	223000.0	327000.0	0.0	0.0
4	2008	PUNGGOL	5 ROOM	305000.0	428000.0	0.0	0.0
5	2008	JURONG WEST	3 ROOM	142000.0	160000.0	0.0	0.0
6	2008	JURONG WEST	4 ROOM	211000.0	253000.0	0.0	0.0
7	2008	JURONG WEST	5 ROOM	229000.0	319000.0	0.0	0.0
8	2008	BUKIT PANJANG	2 ROOM	82000.0	106000.0	0.0	0.0
9	2008	BUKIT PANJANG	3 ROOM	138000.0	170000.0	0.0	0.0
10	2008	BUKIT PANJANG	4 ROOM	211000.0	270000.0	0.0	0.0
11	2008	WOODLANDS	3 ROOM	116000.0	164000.0	0.0	0.0
12	2008	WOODLANDS	4 ROOM	184000.0	257000.0	0.0	0.0
13	2008	WOODLANDS	5 ROOM	247000.0	296000.0	0.0	0.0
14	2008	SENGKANG	2 ROOM	73000.0	95000.0	0.0	0.0
15	2008	SENGKANG	3 ROOM	120000.0	162000.0	0.0	0.0
16	2008	SENGKANG	4 ROOM	190000.0	275000.0	0.0	0.0
17	2008	SENGKANG	5 ROOM	290000.0	367000.0	0.0	0.0
18	2009	PUNGGOL	2 ROOM	89000.0	114000.0	0.0	0.0
19	2009	PUNGGOL	3 ROOM	151000.0	188000.0	0.0	0.0

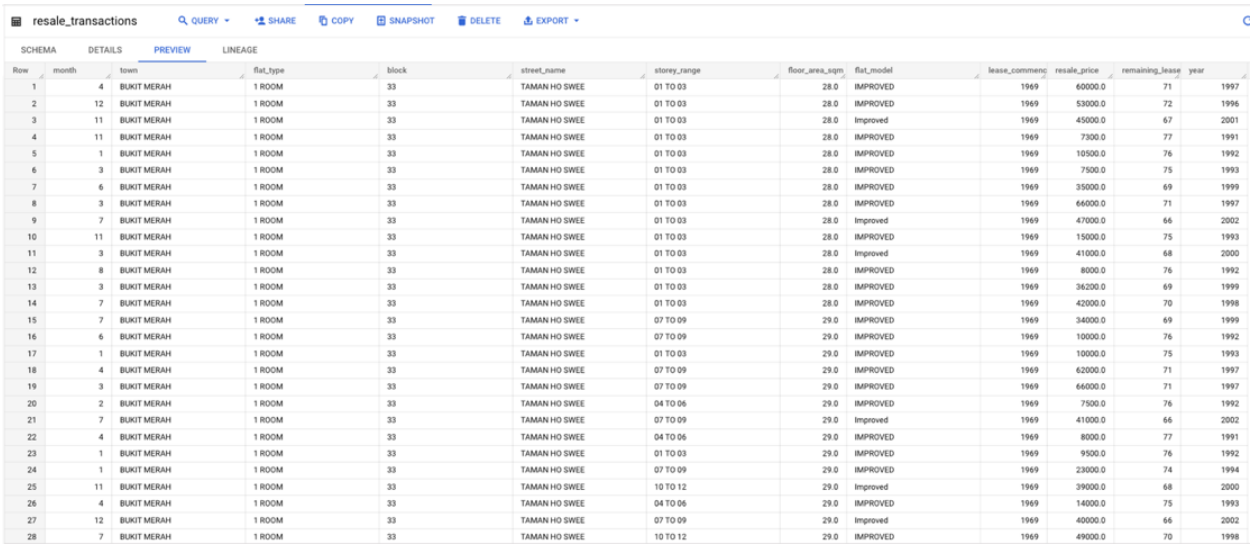
Fig 12. BTO Price Ranges Preview

financial_data QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW LINEAGE

Row	Date	_10Y_Treasury	S_P	STI	month	year
1	1996-01	5.58099985...	636.020019...	2449.19995...	1	1996
2	2015-01	1.67499995...	1994.98999...	3391.19995...	1	2015
3	1990-01	8.43000030...	329.079986...	1515.0	1	1990
4	2002-01	5.02500009...	1130.19995...	1786.89001...	1	2002
5	2001-01	5.17899990...	1366.01000...	1991.29003...	1	2001
6	2009-01	2.84400010...	825.880004...	1746.46997...	1	2009
7	1997-01	6.50299978...	786.159973...	2216.5	1	1997
8	1998-01	5.51200008...	980.280029...	1259.90002...	1	1998
9	2017-01	2.45099997...	2278.87011...	3046.80004...	1	2017
10	1992-01	7.30999994...	408.779998...	1529.69995...	1	1992
11	2014-01	2.66799998...	1782.58996...	3027.21997...	1	2014
12	2023-01	3.52900004...	4076.60009...	3365.66992...	1	2023
13	1999-01	4.65299987...	1279.64001...	1428.14001...	1	1999
14	2022-01	1.78199994...	4515.54980...	3249.59008...	1	2022
15	2021-01	1.09300005...	3714.23999...	2902.52001...	1	2021
16	2012-01	1.79900002...	1312.41003...	2906.68994...	1	2012
17	2006-01	4.52699995...	1280.07995...	2412.08007...	1	2006
18	2016-01	1.93099999...	1940.23999...	2629.11010...	1	2016
19	2005-01	4.13299989...	1181.27001...	2096.32006...	1	2005

Fig 13. Financial Data Preview



resale_transactions

QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW LINEAGE

Row	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commenc	resale_price	remaining_lease	year
1	4	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	60000.0	71	1997
2	12	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	53000.0	72	1996
3	11	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	Improved	1969	45000.0	67	2001
4	11	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	7300.0	77	1991
5	1	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	10500.0	76	1992
6	3	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	7500.0	75	1993
7	6	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	35000.0	69	1999
8	3	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	66000.0	71	1997
9	7	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	Improved	1969	47000.0	66	2002
10	11	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	15000.0	75	1993
11	3	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	Improved	1969	41000.0	68	2000
12	8	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	8000.0	76	1992
13	3	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	36200.0	69	1999
14	7	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	28.0	IMPROVED	1969	42000.0	70	1998
15	7	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	IMPROVED	1969	34000.0	69	1999
16	6	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	IMPROVED	1969	10000.0	76	1992
17	1	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	29.0	IMPROVED	1969	10000.0	75	1993
18	4	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	IMPROVED	1969	62000.0	71	1997
19	3	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	IMPROVED	1969	66000.0	71	1997
20	2	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	04 TO 06	29.0	IMPROVED	1969	7500.0	76	1992
21	7	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	Improved	1969	41000.0	66	2002
22	4	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	04 TO 06	29.0	IMPROVED	1969	8000.0	77	1991
23	1	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	01 TO 03	29.0	IMPROVED	1969	9500.0	76	1992
24	1	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	IMPROVED	1969	23000.0	74	1994
25	11	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	10 TO 12	29.0	Improved	1969	39000.0	68	2000
26	4	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	04 TO 06	29.0	IMPROVED	1969	14000.0	75	1993
27	12	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	07 TO 09	29.0	Improved	1969	40000.0	66	2002
28	7	BUKIT MERAH	1 ROOM	33	TAMAN HO SWEE	10 TO 12	29.0	IMPROVED	1969	49000.0	70	1998

Fig 14. Resale Transactions Preview

Row	comment_id	comment_text	date	polarity_score
1	3693a077fc7fcf0aa49e5628a1...	BTO remains a lottery; still over subscribed, supply lagging behind demand, insane construction times etc. In my humble opinion,	2023-04-17 09:16:08 UTC	-0.0305555...
2	66a3dd0bfe0c418fa7b58e040...	I think the current housing situation is quite different from how it is described in the article.	2023-04-17 08:07:34 UTC	0.08137254...
3	7afad2d458b2ca3e7a49e7848...	History is repeating itself, but this time, only in a more ridiculous way. The housing shortage in 2010-2011 created a spike in	2023-04-17 12:21:59 UTC	0.10849567...
4	444ecdd734673f0283fd8c9be...	Based on the property you are currently renting, waiting could cost you more. Those with unreasonable price (+COV)	2023-04-17 09:17:31 UTC	-0.0349999...
5	adf64f8376e15f2fa169661f68...	My fiancée and I recently	2023-04-17 13:31:12 UTC	0.23970057...

Fig 15. Reddit Comments Preview

Run Times

Task ID	**Run Time**	**Purpose**
yf_update_data	1 second	Update financial data table
delete_rows	5 seconds	Delete old data from resale transactions table
insert_data	13 seconds	Insert new data into resale transactions table
upload_input_data	20 seconds	Generate new input data for ML prediction
delete_combine	4 seconds	Delete old data from combined data table
insert_combine	4 seconds	Insert new data into combined data table
update_model	1 minute 18 seconds	Retrain ML model with new data in combined data table

model_predict	5 seconds	Predict resale prices from input data using new ML model
Total	3 minutes 9 seconds	Update financial data and resale transactions, join new entries, and update combined data. Retrain ML model using new combined data and predict resale prices in the future.

Fig 16. Resale Fin DAG runtimes

Task ID	Run Time	Purpose
delete_rows	4 seconds	Delete old data from BTO price ranges table
Insert_data	2 seconds	Insert new data into BTO price ranges table

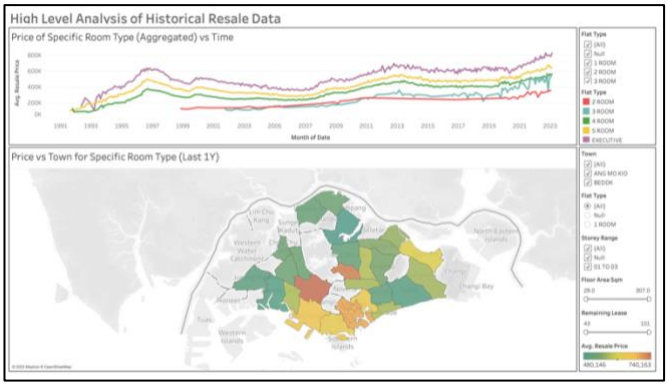
Fig 17. BTO DAG runtimes

Task ID	Run Time	Purpose
get_oauth_token	2 seconds	Get authentication token for Reddit API
get_post_id	4 seconds	Get post ids of most recent 100 posts in r/singapore
get_comments	1 minute 32 seconds	Get comments from the posts
insert_comments_to_bigquery	1 second	Delete old table, insert comments to BigQuery table
Total	2 minutes 7 seconds	Obtain authentication token, retrieve posts ids, get the comments from the post and calculate polarity scores, delete old table and insert new comments in BigQuery.

Fig 18. Reddit DAG runtimes

High Level Analysis of Historical Resale Data

The second page of our dashboard includes filters for Flat Type, Town, Storey Range, Floor Area, Remaining Lease and Resale Price to enable users to better select their ideal property. This page comprises of 2 charts – a line chart to depict the aggregated price of a specific room type across time and a heatmap with the price of the property of a specific room type in each town. Users can better determine a suitable room size that meets their budget and location preferences with this page.



It is noteworthy that 3-room flats have exhibited greater price volatility since the pandemic, with a significantly higher variance in the overall price compared to other types of flats. Therefore, it may be beneficial to closely monitor the supply and demand of 3-room flats to achieve better pricing considering these fluctuations.