Question 1

The HDB Resale Portal was officially launched in Jan 2018 to streamline transactions of HDB resale flats. This has made it easier for buyers and/or sellers to carry out transactions on their own, without utilising a property agent's professional services. This report quantifies the business impact on property agents, using resale transaction data and property agent transaction records.

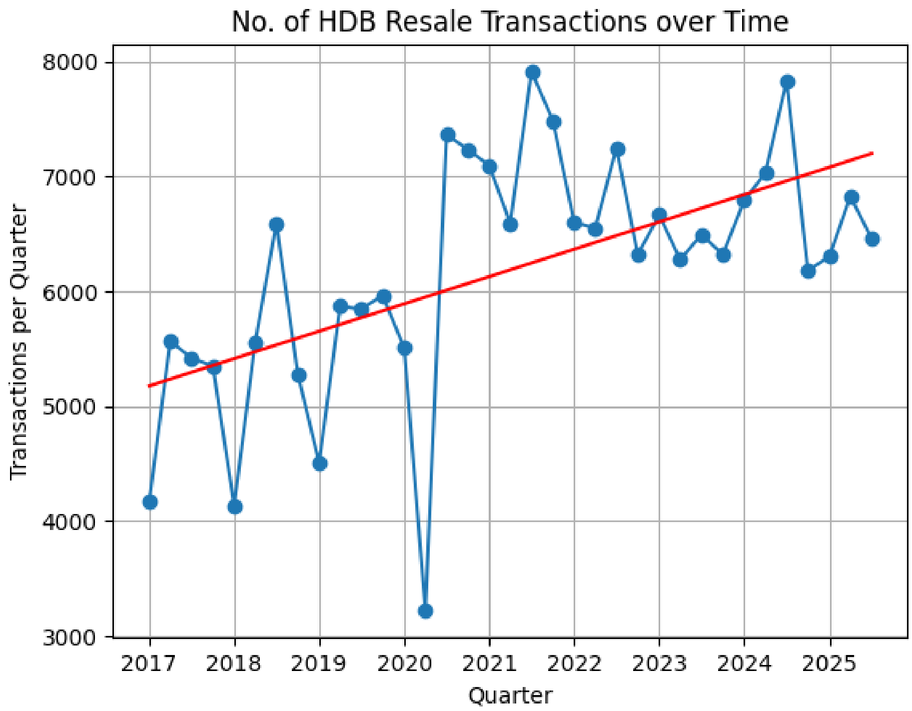

Figure 1: HDB Transaction Volume (2017 – 2025)

Overall Trends in HDB Resale Transactions

Figure 1 shows quarterly HDB resale transaction volumes from 2017 to 2025. Apart from the sudden dip in 2025 (likely due to COVID 19), the trend shows a steady growth from 2017 to 2025, rising from an average of 5200 per quarter in 2017 to around 7000 per quarter in 2025. despite a sudden dip in 2020, likely attributed to COVID-19. Hence, this shows that business impact on agentsshould be interpreted relative to total transaction volumes, not just in raw values.

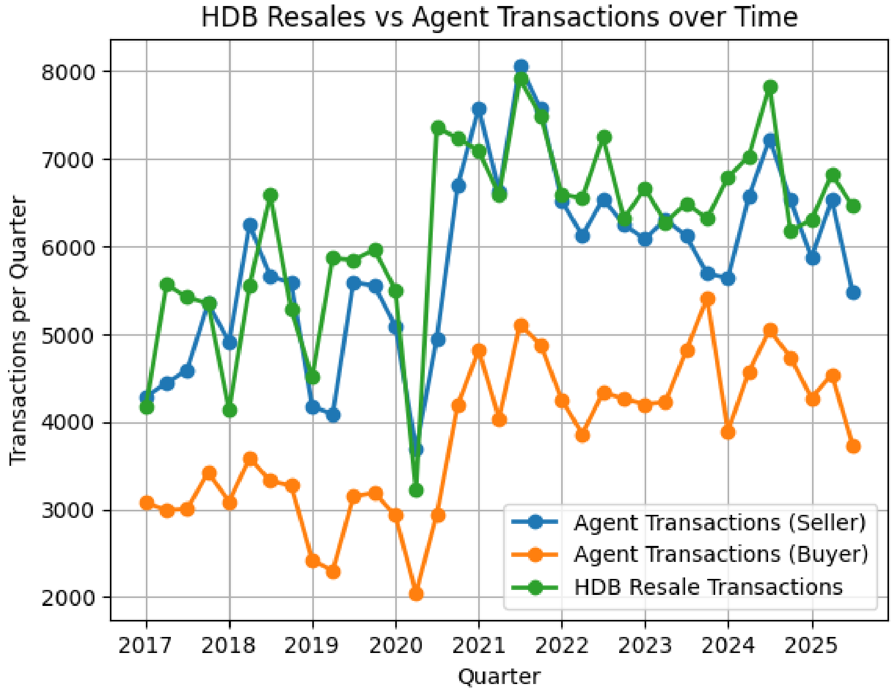

Figure 2: HDB Transaction Volume vs Agent Transactions (2017 – 2025)

Agent Representation (Buyers vs Sellers)

From Figure 2, we compare the agent representation across their roles (representing buyers vs sellers). We can see that the agent transaction for sellers closely follows the total number of HDB resale transactions over time, suggesting that most sellers continue to rely heavily on agents when making transactions, likely due to the need for pricing strategy, marketing and handling paperwork. On the other hand, agent transactions for buyers are consistently lower, implying that buyers find it easier to handle the purchase independently compared to sellers.

Overall, we can see that agent transactions for buyer and seller show an upward trend over time, indicating that agents transactions are still thriving despite the portal being implemented.

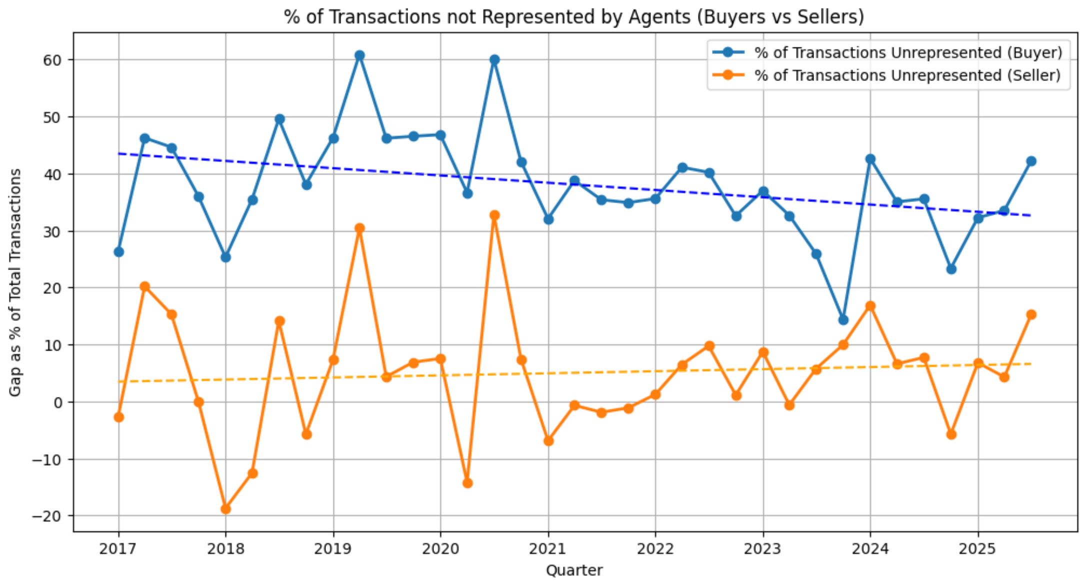

Figure 3: % of Transactions not represented by Agents (Buyers vs Sellers)

Relative Share of Agent Transactions

To isolate the effects of overall transaction growth, we analyse the percentage of transactions (both buyers and sellers) that were not handled by agents, indicating that owners/buyers chose to handle the transaction independently.

From Figure 3, we can see that the % of seller transactions that were not handled by agents remains low from 2017 – 2025, with only a small increase from 4% to 7%. Conversely, the % of buyer transactions that were not handled by agents decreased from 44% in 2017 to 33% in 2025. Rather than being made obsolete by the HDB Resale Portal, which aimed to empower DIY transactions, the share of transactions involving agents for buyers have in fact increased over time.

Conclusion

In conclusion, despite the intent for the HDB Resale Portal to reduce the need for buyers/sellers to engage the help of property agents, the evidence suggest otherwise. Overall, the share of transactions has not declined, but rather shown an upward trend, showing that the services of property agents remain in demand despite the introduction of a digital alternative.

Question 2

Data Preparation

We merge the publicly available datasets between 2011 and 2015, using 'case_number' as the unique identifier key. For this modelling, we take 'type_of_dispute' and 'type_of_intake' as our features, and 'outcome_of_cases' as the label.

After merging, rows (1) with missing values were removed, and categorical features were encoded using OneHotEncoding. We then apply Random Oversampler as a balancing method in order to handle skewed class distributions, particularly in the minority class of "Mediation Without Settlement".

Modelling

For the modelling approach, we break it down into a 2-step classification task, each addressed by a XGBoost Model.

Model 1 – Mediation vs Non Mediation

- Predicts whether a case will proceed to mediation
- Model: XGBoost Classifier with hyperparameter tuning, optimized for imbalanced data using Area under Precision-Recall Curve (AUCPR) as the evaluation metric

Model 2 – Without vs Without Settlement

- Predicts the outcome of mediated cases, on whether they end with settlement or not
- Model: XGBoost Classifier with hyperparameter tuning, applying oversampling to reduce class imbalance between "With Settlement" and "Without Settlement" classes.

Evaluation

Both models were trained on a train-test split of 70-30. For each, the accuracy and classification report was printed out, showing the precision, recall and F1-score per class

Model 1 – Mediation vs Non Mediation

```
--- Model 1: Mediation vs Not ---
Accuracy: 0.7257989810097267
              precision    recall  f1-score   support

    Mediated       0.64      0.67      0.65       833
Not Mediated       0.79      0.76      0.77      1326

    accuracy                           0.73      2159
   macro avg       0.71      0.72      0.71      2159
weighted avg       0.73      0.73      0.73      2159
```

Model 1 achieved an overall accuracy of 72.58%, indicating a reasonable performance for the model given the lack of information available about each case. When comparing the class-specific metrics, we note that the model performed better when predicting “Not Mediated” Classes as compared to “Mediated” Classes. Additionally, the low recall of 0.67 for “Mediated” Classes might mean that the model ma sometimes inaccurately misclassify cases that actually proceed to mediation as “Not Mediated”. While Model 1 provides a good starting point, further information would make it more accurate and improve its ability to classify these cases.

Model 2 – With vs Without Settlement

```
--- Model 2: With vs Without Settlement ---
Accuracy: 0.7286914765906363
                    precision    recall  f1-score   support

   With Settlement       0.75      0.95      0.84       621
Without Settlement       0.33      0.07      0.11       212

          accuracy                           0.73       833
         macro avg       0.54      0.51      0.48       833
      weighted avg       0.64      0.73      0.65       833
```

Model 2 achieved an overall accuracy of 72.87%. While it has high recall (0.75) and precision (0.95) for identifying “With Settlement” cases, it struggles with dealing with the minority class of “Without Settlement” even with random oversampling. This is likely because “Without Settlement” consists of 20% of the Mediated Cases, and the distribution of labels for each feature combination does not have an obvious pattern as well. Hence, while the model is reliable at flagging out cases that will likely result in a settlement, it misses most cases where mediation does not reach a settlement.

This suggests that additional techniques, such as more aggressive resampling may be needed to improve detection of the “Without Settlement” group. However, this may inadvertently result in overfitting of the model, and we believe that a better way to go about improving this model would be to obtain further information for each case, which would allow the model to predict more accurately apart from the 2 features that we currently have.

Question 3

Pipeline

1. **Text and Information Extraction**
   Convert the syllabus and exam paper pdfs into plain text using PyPDF2
   Scan through the exam paper to extract questions and mark allocation, this can be done through regex, looking for question number or mark allocation formats in the text
2. **AI Analysis**
   For each question, we send it along with the syllabus text into a language model, with a well-defined prompt to obtain insights on each question.
   The AI is asked if the question aligns with syllabus objectives, the topics the question covers, and whether the mark allocation fits the weightage for that topic. The prompt can then be structure such that clean and structured responses are returned for each question in a .json file
3. **Overall AI Analysis**
   After getting the breakdown from the AI Analysis of each question, all the information for each individual question can be passed into a language model again, to aggregate the results and return an overall analysis of the entire paper.

Evaluation

The pipeline makes the process transparent and repeatable, as the workflow is already set beforehand, and the process will remain the same regardless of the papers/syllabus that are being passed in. Additionally, mark distribution compared to topic weightage in the syllabus is a quantitative check, which LLMs can handle reliably.

However, the current pipeline would require question papers to be in a specific format, or would require more effort to collate more regex expressions such that questions and mark breakdowns can be reliably extracted. Additionally, this pipeline is highly dependent on accurate and well explained syllabus. If the syllabus contains vague descriptions of topics or mark distribution, the AI may have difficulties verifying the mark distribution.

Generalisation

This pipeline will work well for STEM subjects, where topics are more well-defined (e.g. Algebra, Nervous System, etc.), making it easier for the language model to categorise. In the case of humanities or topics with more broad topics, the LLM may struggle to detect clear topics