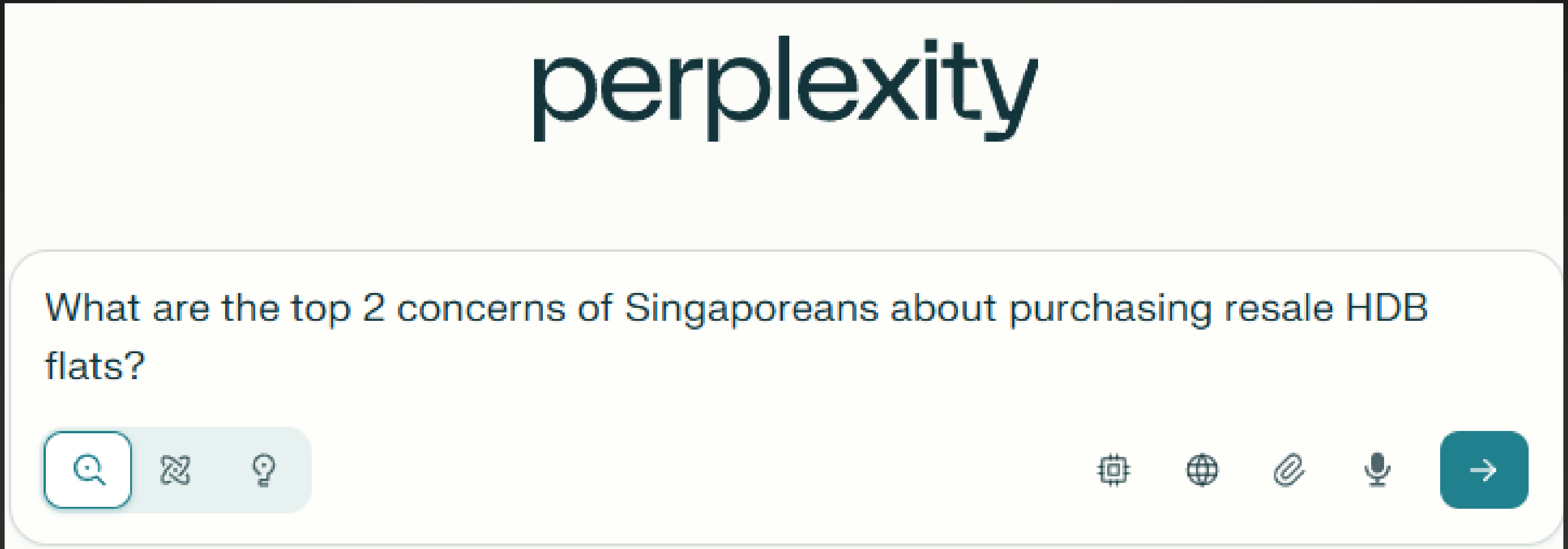
A HDB RESALE ETL PIPELINE PROJECT

~ TAMING THE MARKET, ONE DATASET AT A TIME

RESALE WRANGLES

Azfar, Kent, Mashure, Rakhi



What are the top 2 concerns of Singaporeans about purchasing resale HDB flats?

1. Lease Decay and Future Value

Many are worried about the shortening 99-year lease. As a flat gets older, its value declines, and buyers may face difficulty selling it or unlocking its equity for their retirement.

2. Affordability and Rising Prices

The rise in million-dollar flats highlights this issue, putting financial pressure on buyers and potentially affecting their long-term retirement savings.

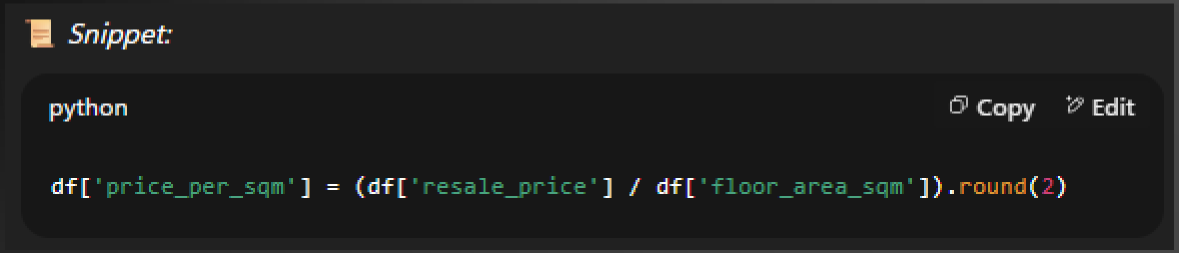
OUR DATA ACQUISITION

- **Source**: Data from the **data.gov.sg API**.
- **Issue**: The full dataset (1990-2025) was not in a single file.
- **Solution**:
  - Fetched **five separate datasets.**
  - Consolidated them into a single, complete DataFrame for analysis.
- **Method**: Used **asynchronous requests** to fetch all data pages efficiently and quickly.

WHY SCHEMA MATTERS BEFORE ETL

Why Schema Matters Before ETL

- ETL without schema = chaos (like packing without a list)
- Schema is the blueprint – defines columns, types, relationships
- Prevents wasted effort & inconsistent data



```python
df['price_per_sqm'] = (df['resale_price'] / df['floor_area_sqm']).round(2)
```

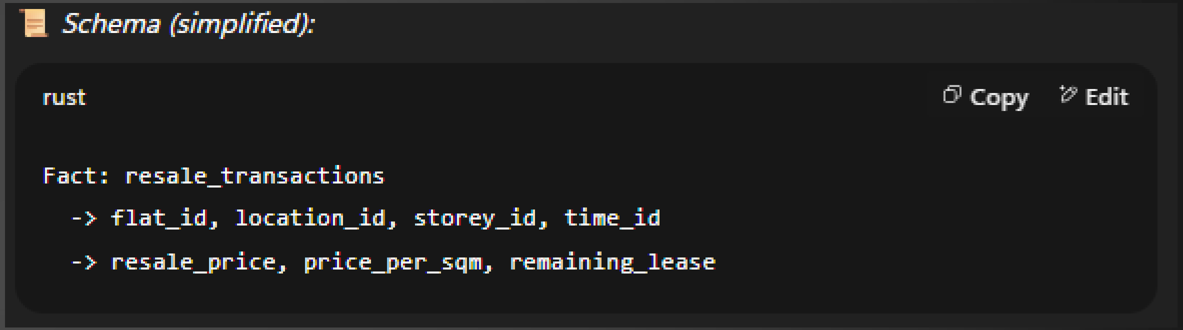
Role of Schema in Analytics

- Ensures consistency in naming & types
- Enables reliable transformations (e.g. price_per_sqm)
- Provides reproducibility & trust in results

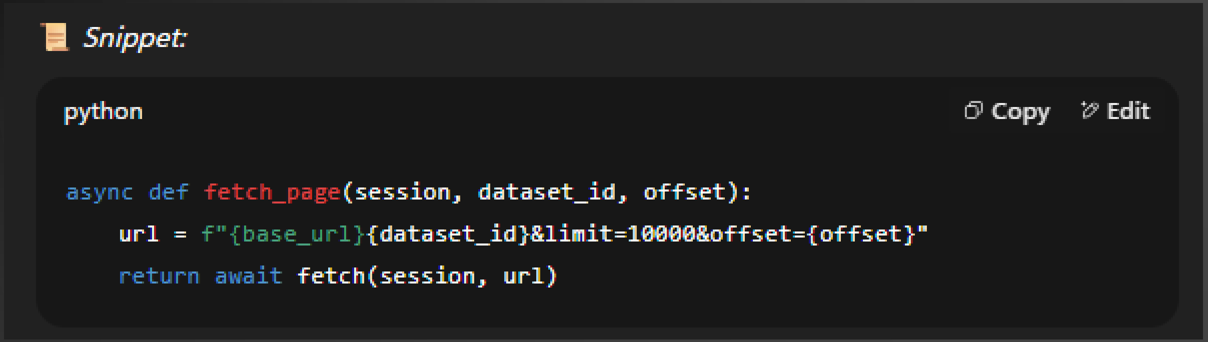
WHY STAR SCHEMA WORKS

Why Star Schema Works Best

- **Fact table**: transactions (price, sqm, lease)
- **Dimension tables**: flat, location, storey, time
- Benefits: fast queries, simple SQL, scalable



```rust
Fact: resale_transactions
  -> flat_id, location_id, storey_id, time_id
  -> resale_price, price_per_sqm, remaining_lease
```



```python
async def fetch_page(session, dataset_id, offset):
    url = f"{base_url}{dataset_id}&limit=10000&offset={offset}"
    return await fetch(session, url)
```

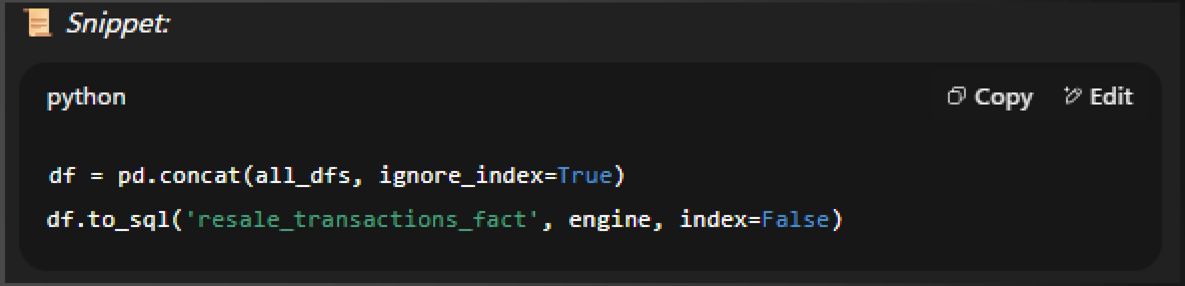
Factors Before Writing a Crawler

- Source reliability: API vs website
- Data volume: async vs slow fetch
- Schema alignment with DB
- Error handling & retries
- Ethics: respect robots.txt

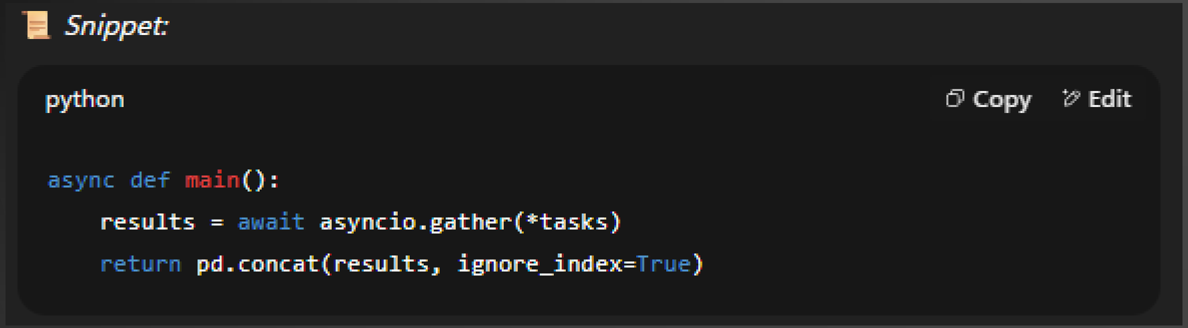
PLANNING CRAWLERS & BENEFITS OF SCRIPTS

Example: HDB Resale Data Pipeline

- 5 datasets fetched from data.gov.sg
- Async crawler using aiohttp + asyncio
- Cleaned with pandas into standard schema
- Loaded into PostgreSQL as star schema



```python
df = pd.concat(all_dfs, ignore_index=True)
df.to_sql('resale_transactions_fact', engine, index=False)
```



```python
async def main():
    results = await asyncio.gather(*tasks)
    return pd.concat(results, ignore_index=True)
```

Advantages of a Well-Written Script

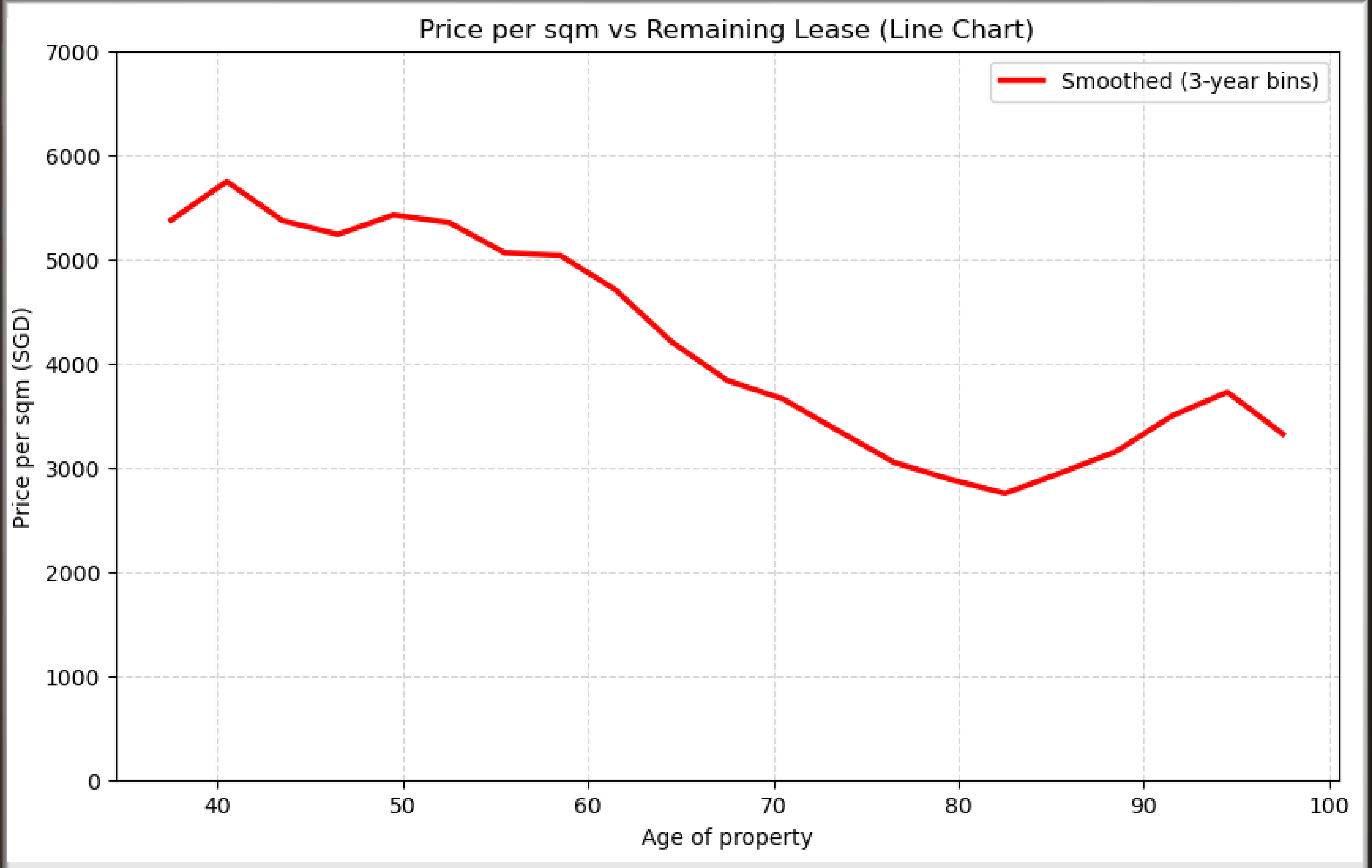
- Speed: async fetching saves hours
- Consistency: same schema every run
- Scalability: add new datasets easily
- Integration: ETL end-to-end automation

FROM QUESTIONS TO DATA

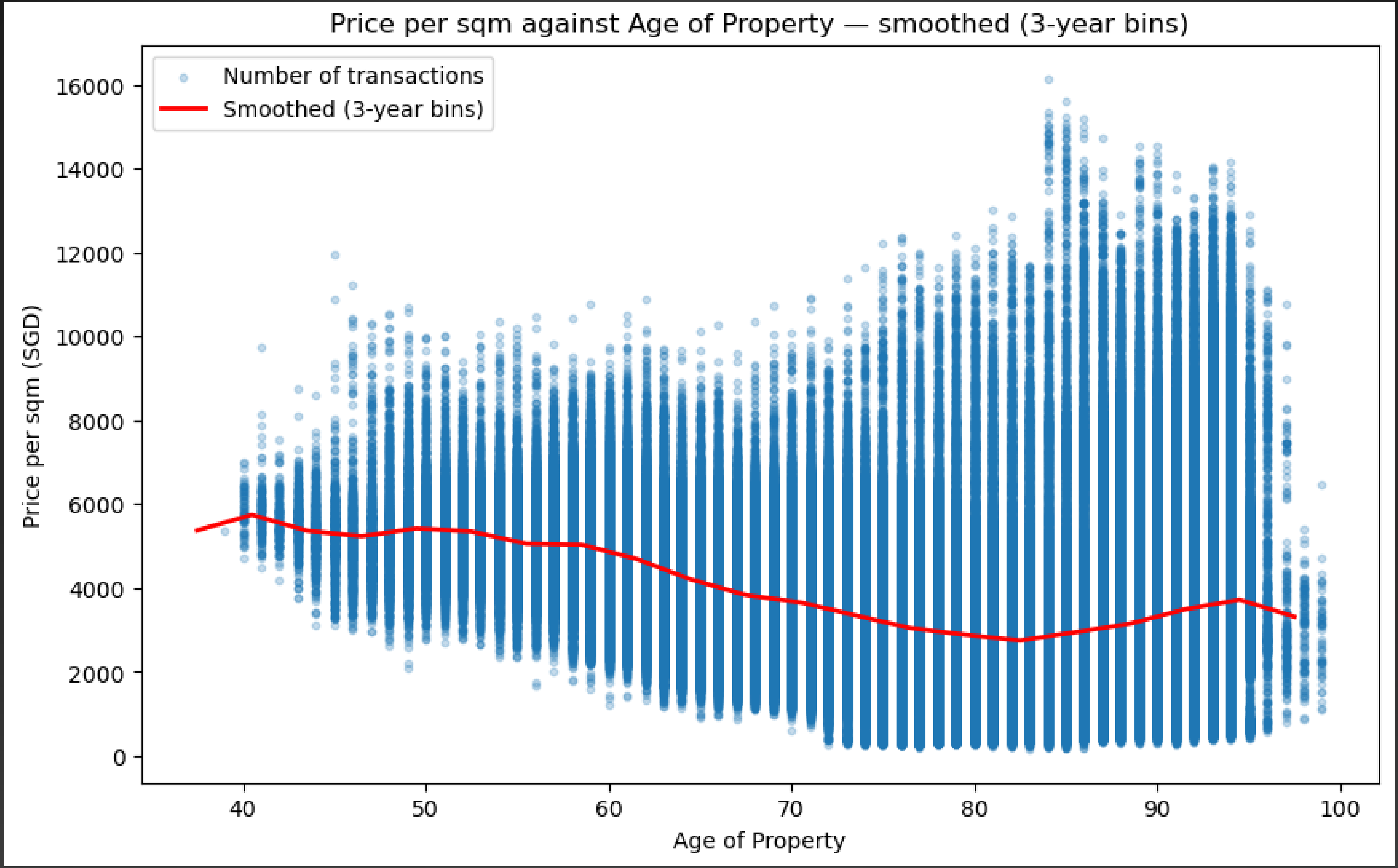
- **We started by asking:** How do concerns like affordability and lease decay affect Singapore's resale flat market?
- **Our approach:** We used a web crawler to gather a complete and clean dataset from a government API to find the answers.
- **Our result:** We successfully built a comprehensive dataset spanning from 1990 to the present, ready for analysis.

1. LEASE DECAY AND FUTURE VALUE

1. LEASE DECAY AND FUTURE VALUE

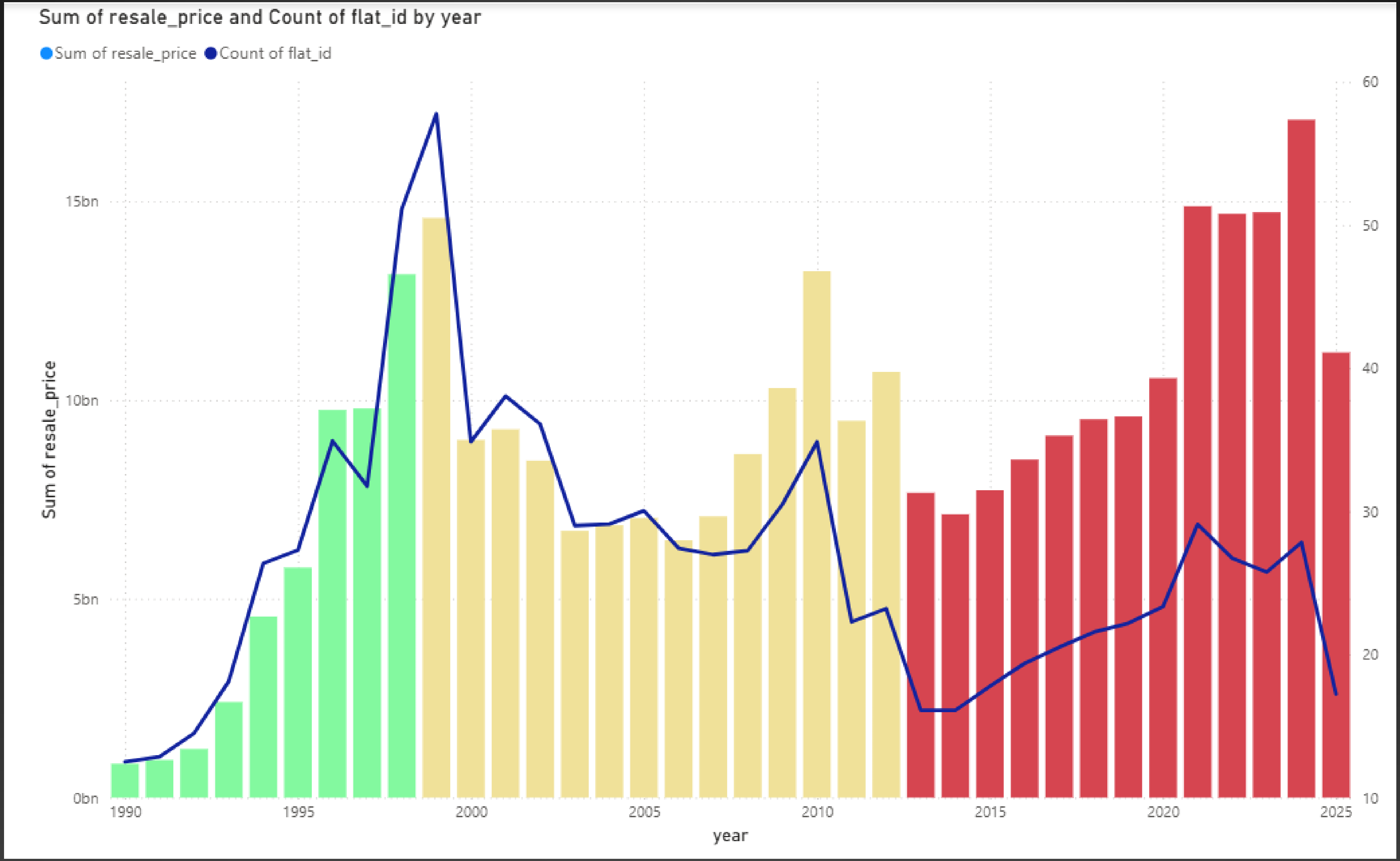


1. LEASE DECAY AND FUTURE VALUE



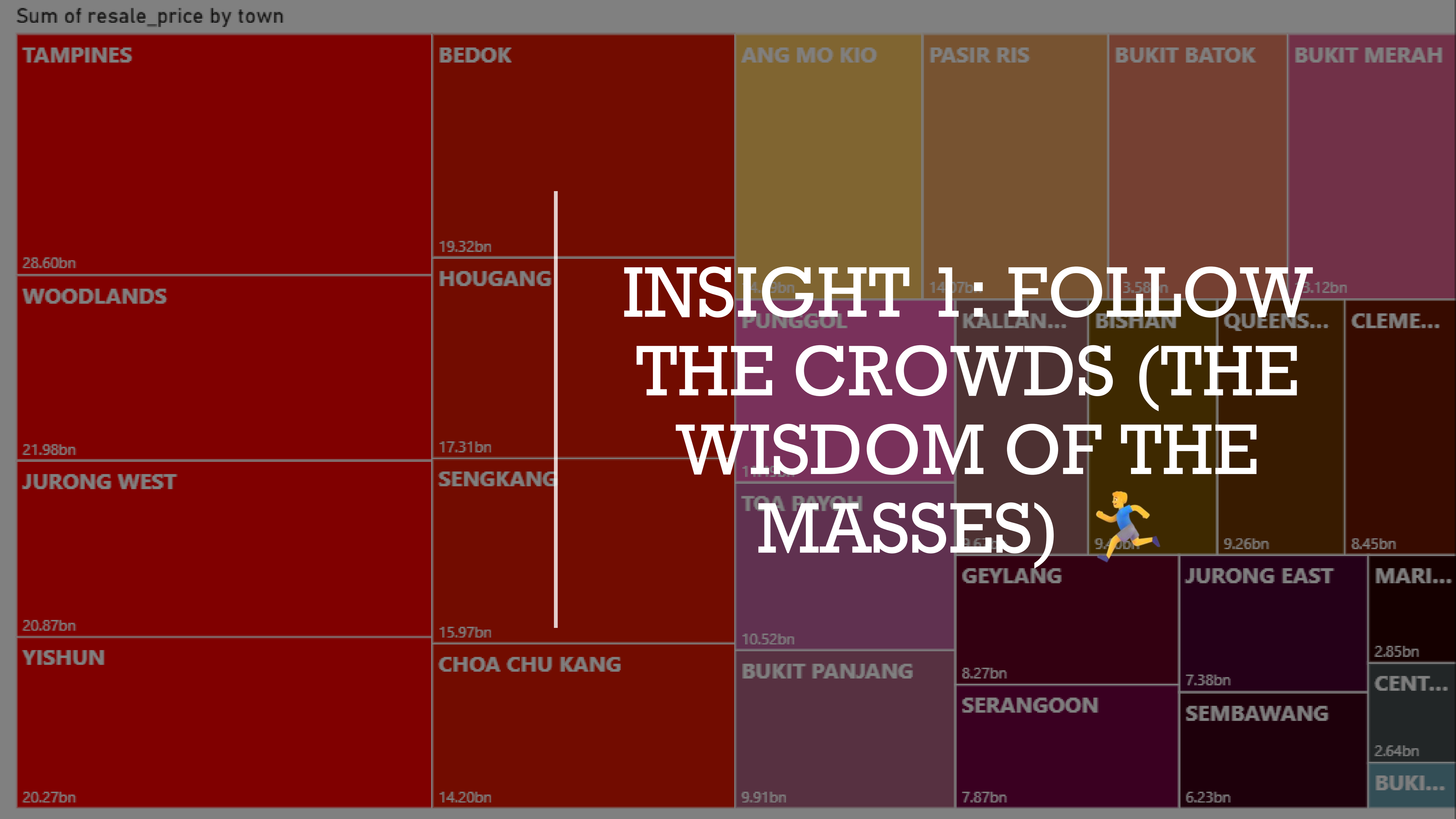
2. AFFORDABILITY AND RISING PRICES

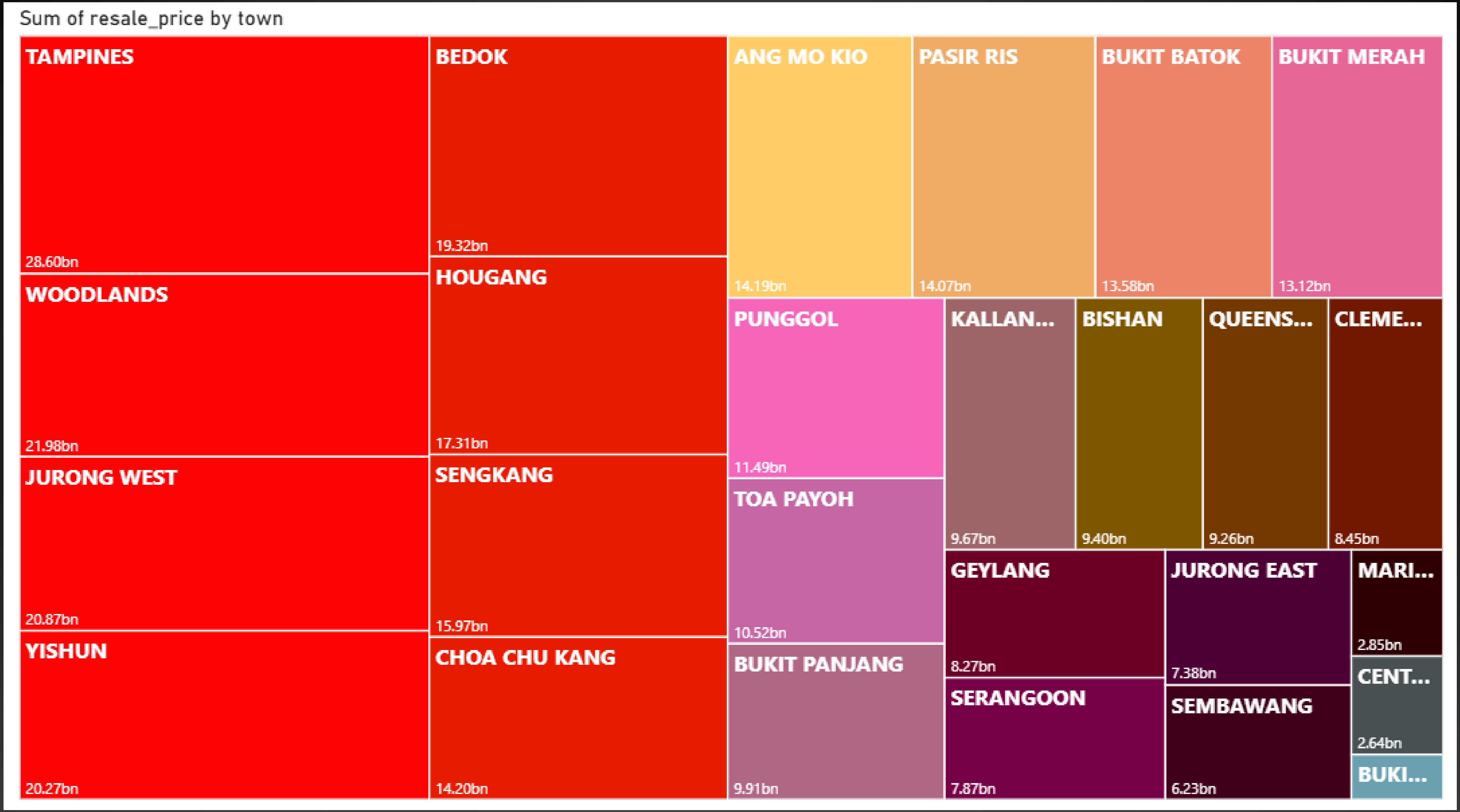
2. AFFORDABILITY AND RISING PRICES

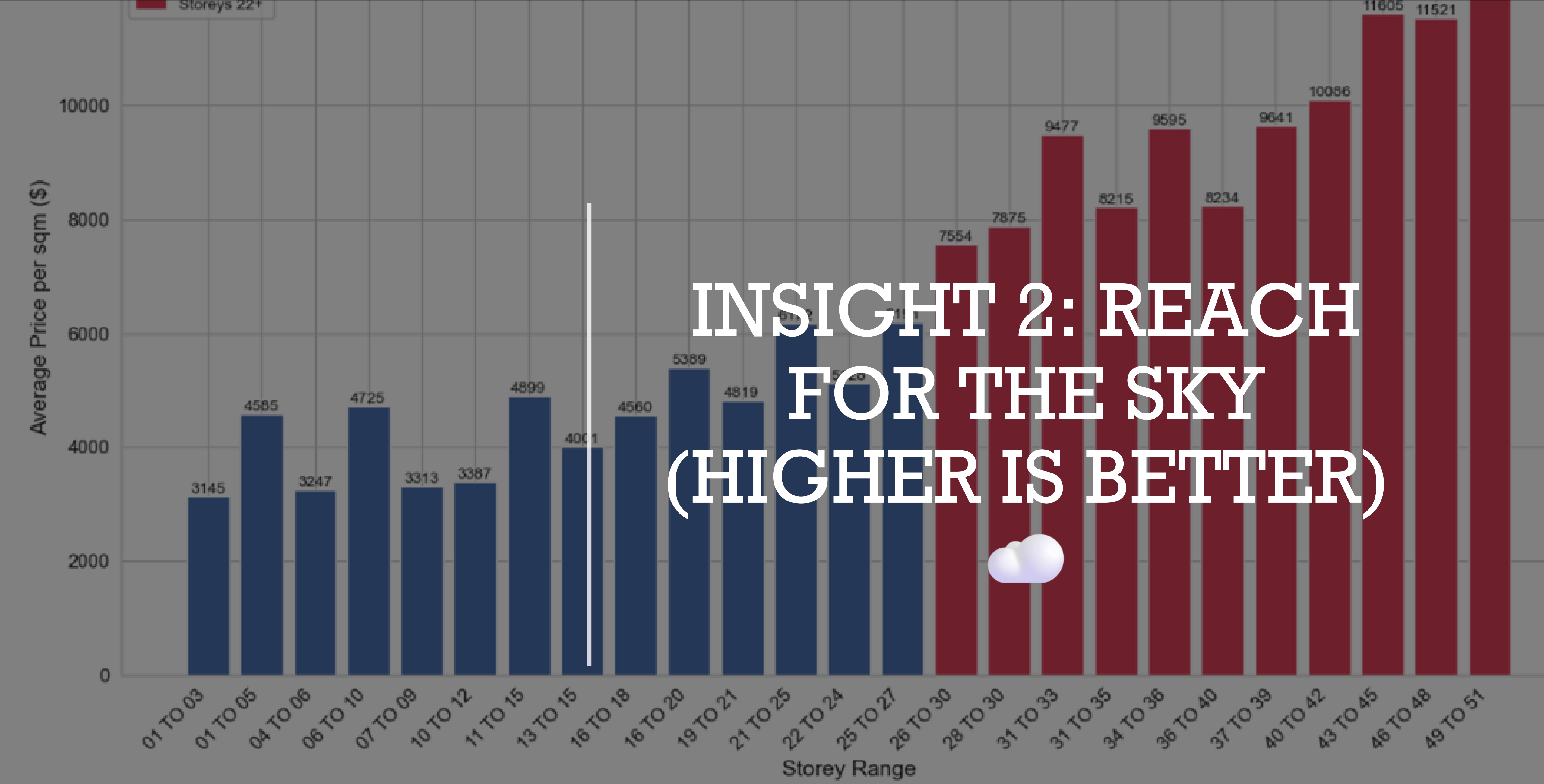


FINAL INSIGHTS: YOUR HDB BUYING CHEAT SHEET

INSIGHT 1: FOLLOW THE CROWDS (THE WISDOM OF THE MASSES) 🏃



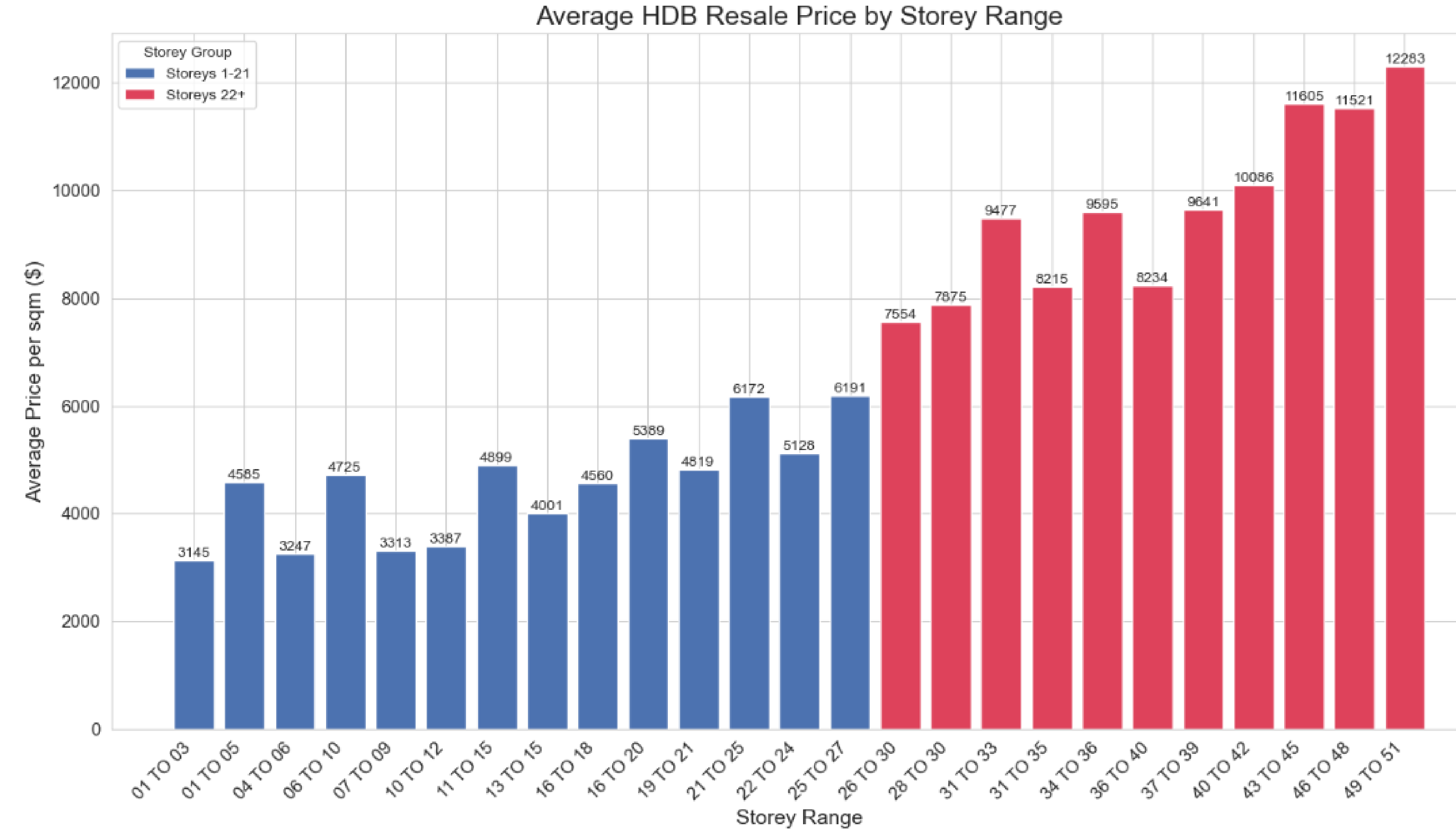

Sum of resale_price by town
TAMPINES
28.60bn
WOODLANDS
21.98bn
JURONG WEST
20.87bn
YISHUN
20.27bn
BEDOK
19.32bn
HOUGANG
17.31bn
SENGKANG
15.97bn
CHOA CHU KANG
14.20bn
ANG MO KIO
14.19bn
PASIR RIS
14.07bn
BUKIT BATOK
13.58bn
BUKIT MERAH
13.12bn
PUNGGOL
11.49bn
TOA PAYOH
10.52bn
BUKIT PANJANG
9.91bn
KALLAN...
9.67bn
BISHAN
9.40bn
QUEENS...
9.26bn
CLEME...
8.45bn
GEYLANG
8.27bn
SERANGOON
7.87bn
JURONG EAST
7.38bn
SEMBAWANG
6.23bn
MARI...
2.85bn
CENT...
2.64bn
BUKI...



INSIGHT 2: REACH FOR THE SKY (HIGHER IS BETTER)

☁️

Don't just invest in a home, invest in a good "storey". Our analysis shows that once you get



Don't just invest in a home, invest in a good "storey". Our analysis shows that once you get above the 28th storey, your property's value really begins to take off. 🚀

THE END