Singapore Birth Rate & Fertility Analysis

GA Data Analytics Immersive – Capstone Technical Report - Lester

01 – Problem Statement, Objectives & Audience

Singapore's Total Fertility Rate (TFR) and resident live births have declined steadily over the past several decades. This long-term decline raises concerns related to population ageing, labour force sustainability, and long-term economic stability.

This analysis examined historical fertility and birth data to understand the scale, timing, and structure of this decline. In addition, predictive modelling was used to assess whether short-term fertility changes could be forecast using historical data.

The objectives of this project were to analyse long-term trends in fertility and births, identify structural patterns over time, and evaluate the performance of simple predictive models based on past fertility behaviour.

This technical report was written for policymakers and planners involved in population policy, workforce development, and long-term socio-economic planning in Singapore.

02 – Data Sources & Data Understanding

The analysis used official Singapore government statistics downloaded in CSV format. Two main datasets were used: Total Fertility Rate (TFR) by year and resident live births by year and birth order.

These datasets were selected because they are directly related to fertility behaviour and birth outcomes, cover long historical periods, and are suitable for time-series analysis. Using official data ensured consistency and reliability across the full time span analysed.

Both datasets provided national-level annual data. While more granular demographic data would allow deeper analysis, national-level data was sufficient to examine long-term structural trends in fertility and births.

In addition, the HDB Resale Price Index was incorporated as a contextual economic indicator in extended modelling to assess its potential explanatory value in relation to fertility trends.

03 – Data Preparation & Cleaning

Before analysis, both datasets were validated to ensure they were accurate, consistent, and suitable for time-series analysis and modelling. Data preparation focused on structural checks, data type verification, missing and duplicate value checks, value range validation, and reshaping.

Structural inspection confirmed that the TFR dataset contained yearly values from 1960 to 2024, stored as numeric values, with no missing or duplicate entries. The births dataset contained yearly birth counts by birth order from 1990 to 2024, also with no missing or duplicate values.

Minimum and maximum value checks showed that TFR values ranged from 0.97 to 5.76 and birth counts ranged from 472 to 49,787. These values were consistent with historical fertility patterns in Singapore and did not indicate invalid data.

Both datasets were originally stored in wide format, with each year represented as a separate column. For analysis and modelling, the data was reshaped into long format using `pandas.melt()`. This structure allowed the data to be sorted chronologically, merged across datasets, and used for time-series analysis.

After reshaping, data types were revalidated and no missing values were introduced. Cleaned datasets and corresponding data dictionaries were saved to document variable definitions, data types, and units of measurement.

In [97]:
```
print("=== Total Fertility Rate Data Dictionary ===")
display(tfr_dictionary)

print("=== Births Data Dictionary ===")
display(births_dictionary)

print("=== HDB Resale Price Index Data Dictionary ===")
display(hdb_dictionary)
```

```
=== Total Fertility Rate Data Dictionary ===
```

	Column Name	Data Type	Description	Unit of Measurement
0	Data Series	object	Fertility indicator name	Text
1	Year	int64	Calendar year	Year
2	Total_Fertility_Rate	float64	Total Fertility Rate (TFR)	Live births per female

```
=== Births Data Dictionary ===
```

	Column Name	Data Type	Description	Unit of Measurement
0	Data Series	object	Birth order category	Text
1	Year	int64	Calendar year	Year
2	Birth_Count	int64	Number of resident live births	Count

```
=== HDB Resale Price Index Data Dictionary ===
```

	Column Name	Data Type	Description	Unit of Measurement
0	Year	int64	Calendar year	Year
1	HDB_Resale_Index	float64	Annual average HDB Resale Price Index	Index (Base 100)

04 – Exploratory Data Analysis (EDA)

EDA

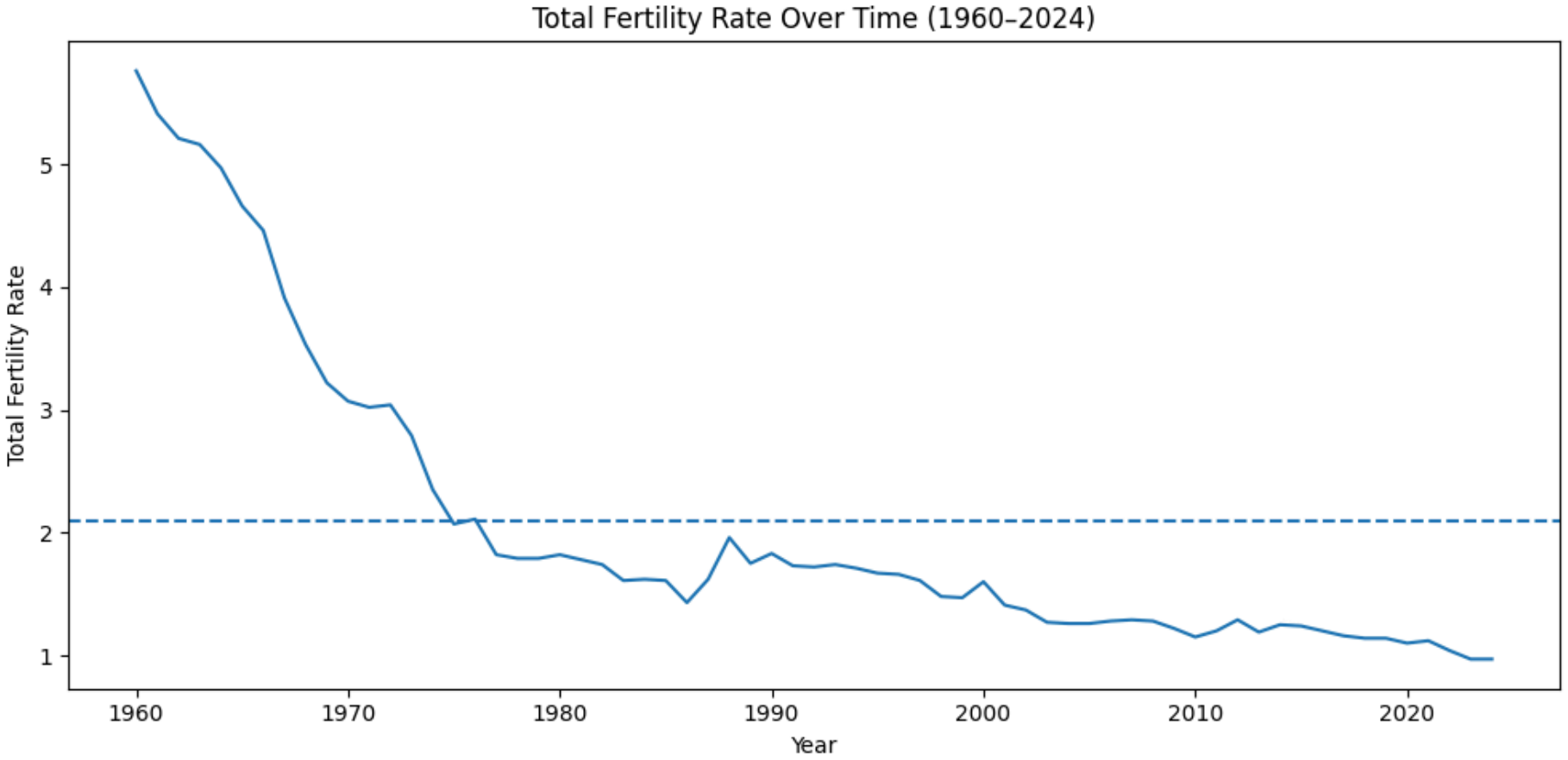
Exploratory analysis was conducted to examine historical trends in Total Fertility Rate and resident live births. The aim was to describe long-term fertility patterns, changes in birth volumes, shifts in birth order composition, and the relationship between fertility and birth counts.

Total Fertility Rate peaked at 5.76 in 1960 and declined to 0.97 by 2024, representing an overall decline of approximately 83%. The steepest decline occurred between 1960 and the mid-1970s, followed by a slower but persistent decline in later decades.

In [98]:
```
plt.figure(figsize=(10, 5))
plt.plot(tfr_long["Year"], tfr_long["Total_Fertility_Rate"])
plt.axhline(2.1, linestyle="--")
```

```
plt.xlabel("Year")
plt.ylabel("Total Fertility Rate")
plt.title("Total Fertility Rate Over Time (1960–2024)")

plt.tight_layout()
plt.show()
```



A replacement level of 2.1 was used as a reference threshold. TFR first fell below this level in 1975 and remained below replacement for the rest of the observed period. This indicates that population growth has not been sustained by births alone for several decades.

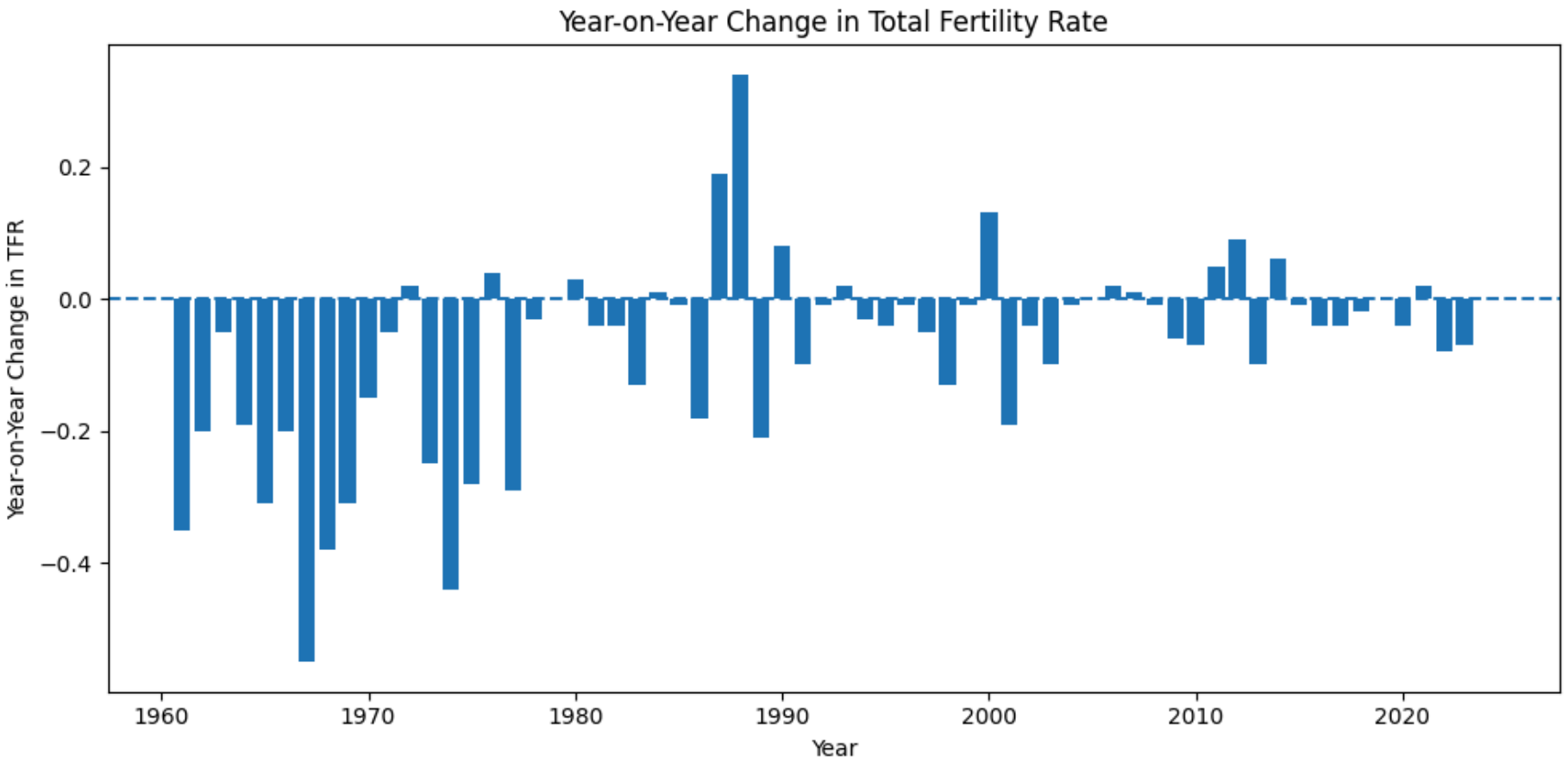
Year-on-year changes in TFR were calculated to assess the pace of decline. The average annual change across the full period was −0.075. When the period was split, the decline was much steeper between 1960 and 1980 than between 2000 and 2024, indicating that most of the sharp fertility decline occurred earlier, followed by a long period of persistently low fertility.

In [99]:
```
tfr_long["TFR_YoY"] = tfr_long["Total_Fertility_Rate"].diff()

plt.figure(figsize=(10, 5))
plt.bar(tfr_long["Year"], tfr_long["TFR_YoY"])
plt.axhline(0, linestyle="--")

plt.xlabel("Year")
plt.ylabel("Year-on-Year Change in TFR")
plt.title("Year-on-Year Change in Total Fertility Rate")

plt.tight_layout()
plt.show()
```



Total resident live births peaked at 99,570 in 1990 and declined to 61,031 by 2023, a decrease of approximately 39%. While short-term fluctuations were observed, the overall trend in birth volumes was downward.

Birth order analysis showed a structural shift in family size patterns. Over time, first-order births accounted for a larger share of total births, while higher-order births declined. This suggests that smaller family sizes have become more common.

In [100…

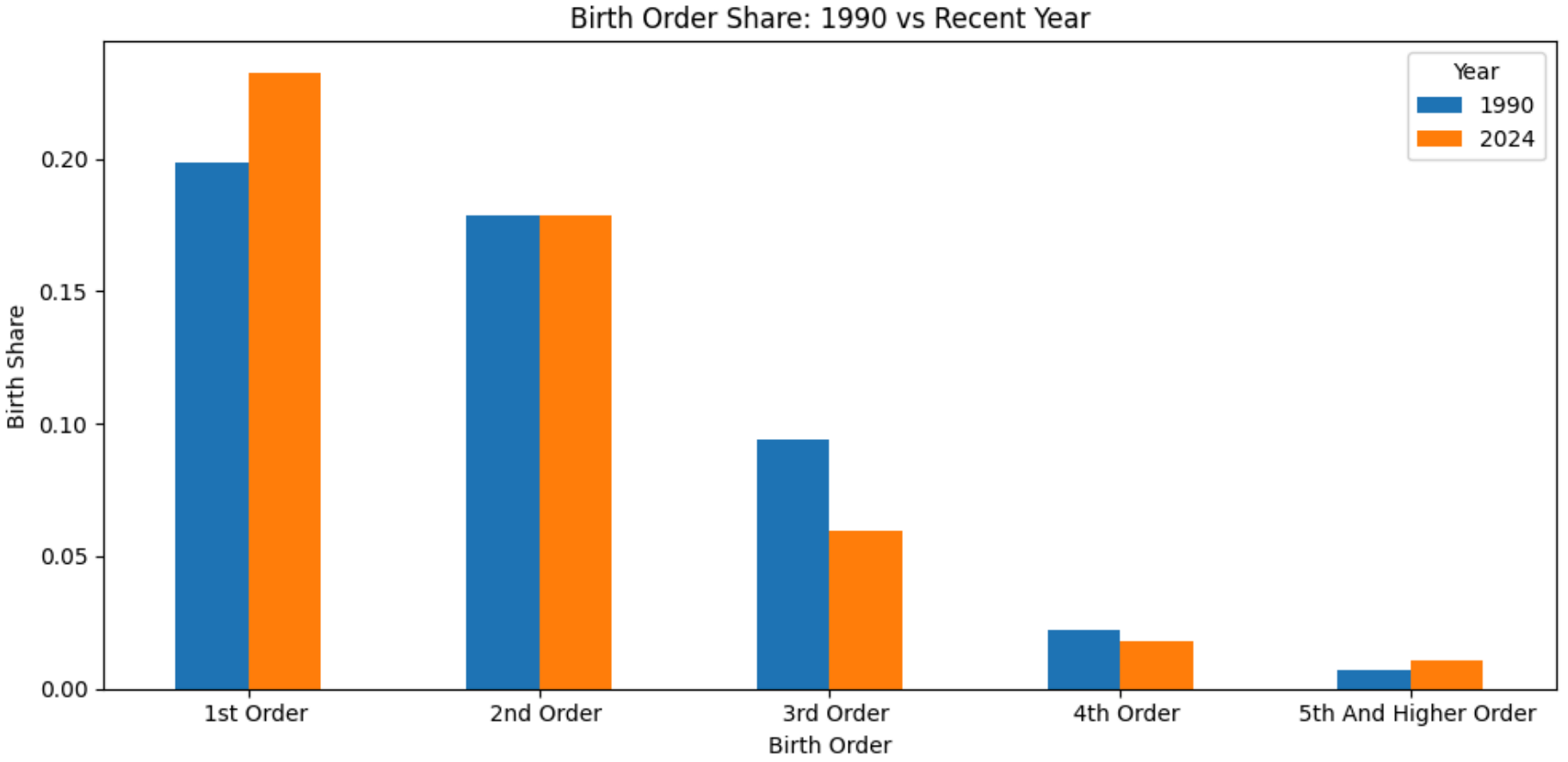
```
# Select two comparison years
comparison_years = births_with_total[
    births_with_total["Year"].isin([1990, births_with_total["Year"].max()])
]

# Pivot for bar chart
```

```
comparison_pivot = comparison_years.pivot(
    index="Data Series", columns="Year", values="Birth_Share"
)

# Plot bar chart
comparison_pivot.plot(kind="bar", figsize=(10, 5))

plt.xlabel("Birth Order")
plt.ylabel("Birth Share")
plt.title("Birth Order Share: 1990 vs Recent Year")
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



Correlation analysis between TFR and total resident live births showed a very strong positive relationship. This relationship was expected because TFR is derived from live birth data. The correlation reflects how the measures are constructed rather than a causal relationship between fertility behaviour and birth volume.

Overall, the EDA indicated a long-term shift towards persistently low fertility, declining birth volumes, and smaller family sizes. These findings provided context for feature engineering and predictive modelling.

05 – Feature Engineering

Feature engineering focused on preparing the Total Fertility Rate dataset for predictive modelling. Given the annual frequency of the data, features were designed to capture temporal persistence, short-term changes, and long-term trends.

A sequential time index was created to represent overall movement across years. Lagged TFR variables from the previous one, two, and three years were included to capture fertility persistence, as fertility behaviour tends to change gradually rather than abruptly. Longer lag structures were not used due to the limited size of the dataset.

A year-on-year change variable was also created to capture short-term fluctuations between consecutive years. Rows affected by lag-related missing values were removed to produce a clean modelling dataset.

Modelling focused on TFR rather than birth counts, as TFR reflects fertility behaviour directly, while birth counts are influenced by population size and composition.

06 – Predictive Modelling

Predictive modelling was conducted to assess whether historical fertility patterns could be used to forecast short-term changes in Total Fertility Rate. Model performance was evaluated relative to a naïve lag-1 baseline, which served as a minimum benchmark.

A time-aware train–test split was used to preserve chronological order. The final 10 years of data were reserved as the test set, with the remaining observations used for training. This approach reflected real-world forecasting conditions and avoided information leakage.

The naïve baseline model used the previous year’s TFR as the prediction for the current year. This model produced low prediction errors and a relatively high R^2, indicating strong year-to-year persistence in fertility.

A linear regression model using lagged TFR values and a time index was then fitted. This model produced higher errors and lower explanatory power than the baseline, indicating that learning a linear relationship did not improve short-term prediction beyond simply using the previous year's value.

To explore whether economic conditions influenced fertility, the HDB Resale Price Index was added as an external variable. Quarterly HDB values were converted to annual averages to align with the TFR data, and the modelling dataset was restricted to overlapping years.

The extended model performed substantially worse than both the baseline and lag regression models. The HDB index was highly correlated with the time index, leading to multicollinearity and unstable predictions.

07 – Model Evaluation & Interpretation

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 on the same test period for all models.

The naïve lag-1 baseline consistently outperformed the regression-based models. Increased model complexity did not improve predictive accuracy, and the inclusion of the HDB Resale Price Index reduced performance further.

These results indicate that short-term fertility changes are dominated by persistence. Past TFR values explain most short-term variation, while additional predictors do not add meaningful predictive value within this framework.

In [101…

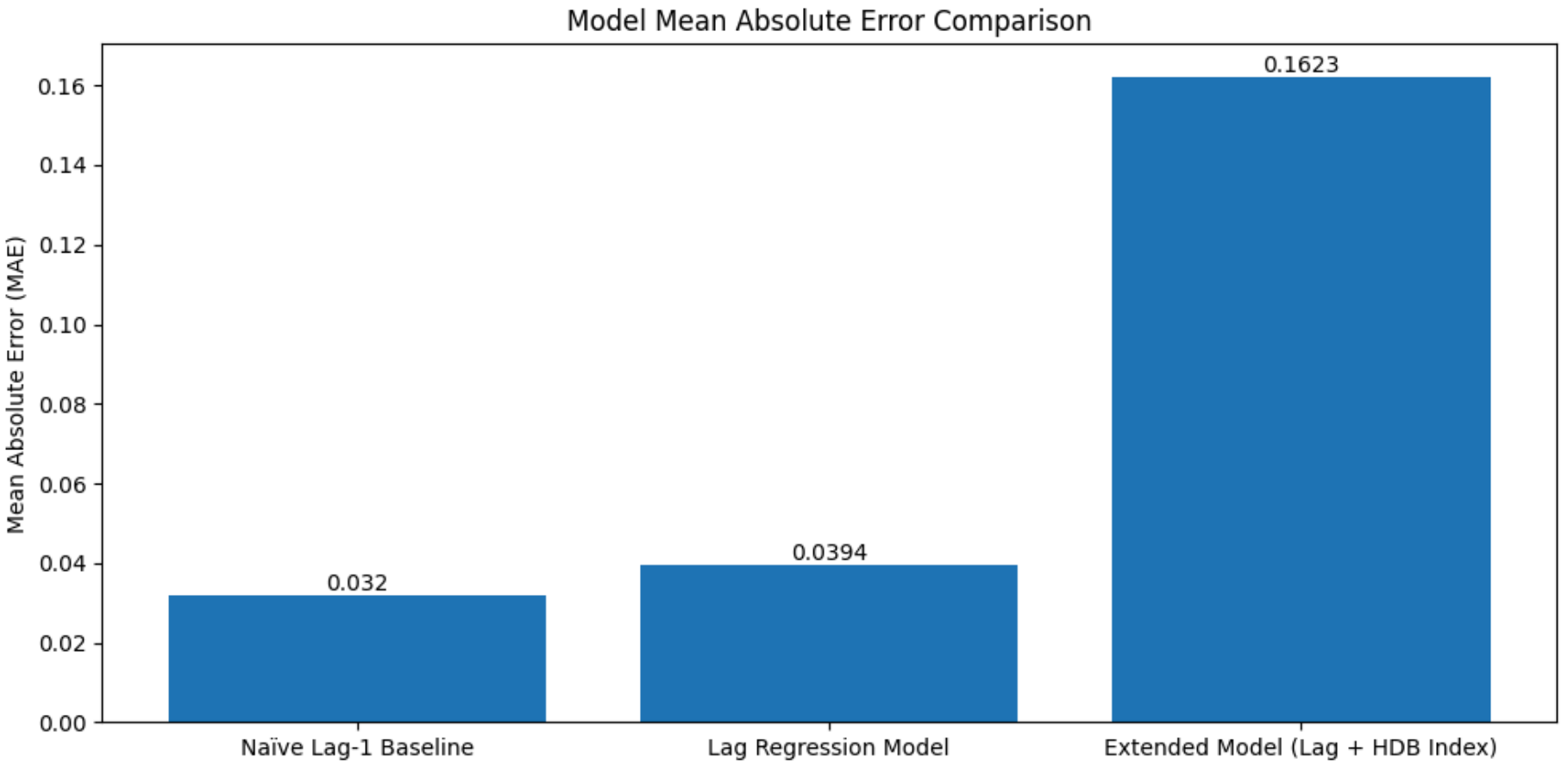
```
plt.figure(figsize=(10, 5))

bars = plt.bar(model_comparison["Model"], model_comparison["MAE"])

for bar in bars:
    h = bar.get_height()
    plt.text(
        bar.get_x() + bar.get_width() / 2, h, round(h, 4), ha="center", va="bottom"
    )

plt.ylabel("Mean Absolute Error (MAE)")
plt.title("Model Mean Absolute Error Comparison")
```

```
plt.tight_layout()
plt.show()
```



08 – Assumptions & Limitations

The analysis assumed that fertility changes gradually over time and that annual data is sufficient to capture long-term fertility trends. Linear regression assumptions were applied, including approximate linearity and independence of errors.

After feature engineering, the modelling dataset contained 62 annual observations. This limited model complexity and increased sensitivity to noise. The use of national-level data may also mask variation across age groups, cohorts, or demographic subgroups.

Policy effects, immigration dynamics, and population age structure were not explicitly modelled. The strong correlation between time and housing prices limited the usefulness of economic variables in simple regression models.

09 – Conclusions & Next Steps

This analysis showed a sustained long-term decline in fertility in Singapore. Total Fertility Rate has remained below replacement level since the mid-1970s, and resident live births have declined alongside a shift towards smaller family sizes.

From a modelling perspective, short-term fertility forecasting was best explained by persistence. A simple lag-1 baseline outperformed more complex regression models, and adding housing price data did not improve predictive performance due to multicollinearity.

Future work could incorporate age-specific or cohort-based fertility data, explicitly model policy timing, or apply alternative time-series approaches designed for persistent behaviour. More granular or higher-frequency data may provide additional insight beyond what is possible with annual national-level data.