UNDERSTANDING AND FORECASTING MARRIAGE TRENDS:
An Open Data Approach Combining Exploratory Analysis and Machine Learning

MUHAMMAD SUKRI BIN RAMLI
Asia School of Business
Kuala Lumpur, Malaysia
Email: m.binramli@sloan.mit.edu

Abstract

This research investigates marriage trends in Malaysia using open data from the Department of Statistics Malaysia (DOSM), with a focus on understanding historical patterns, forecasting future marriage rates, and exploring potential intervention strategies to mitigate the negative impacts of an aging population and low birth rates. Employing a combined approach of exploratory data analysis and machine learning, we aim to analyze historical marriage patterns and forecast future trends. The exploratory analysis examines marriage data by age, gender, and state, revealing demographic patterns and potential influences of socioeconomic and cultural factors. Building on these insights, we develop and evaluate machine learning models, incorporating socioeconomic and demographic indicators, to forecast marriage rates. Our findings highlight the influence of population density and income on marriage rates, forming complex feedback loops. This research contributes to a comprehensive understanding of marriage dynamics in Malaysia, offering valuable insights for policymakers and researchers interested in marriage trends, their societal implications, and potential interventions to address demographic challenges. The study acknowledges limitations in data availability and model accuracy, particularly for long-term forecasts.

1. Introduction

Malaysia's future demographic landscape is projected to be characterized by an aging society, low fertility rates, and declining birth rates. These trends raise concerns about the country's long-term economic growth, social stability, and the sustainability of social support systems. Marriage, as a fundamental social institution, plays a crucial role in shaping family structures and demographic trends (Klein & White, 2018). Understanding marriage patterns is essential for policymakers and researchers seeking to address social and economic issues related to family well-being, population growth, and social change (Carr & Hudson, 2017). In Malaysia, marriage trends have been evolving in recent decades, influenced by a complex interplay of demographic shifts, socioeconomic changes, and cultural factors (Ali & Peng, 2001). Demographic shifts, such as changing age structure and urbanization, have influenced marriage patterns (Ali & Sivamurugan, 2002). Socioeconomic changes, including increasing female labor force participation and rising education levels, have also played a role (Ali & Peng, 2001). Cultural factors, such as shifting attitudes towards marriage and the influence of religious beliefs, have further contributed to the evolving trends (Thornton & Young-DeMarco, 2001). This research focuses on understanding and forecasting these trends, leveraging the power of open data and advanced analytical techniques. To address these issues, this study investigates historical marriage patterns and utilizes machine learning to forecast future marriage rates in Malaysia.

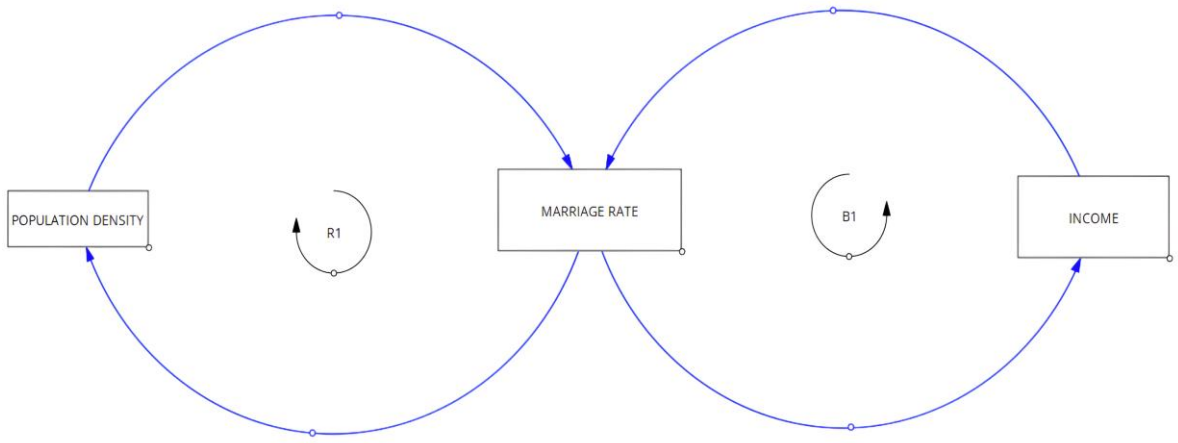

Figure 1: Causal Loop Diagram of Marriage Rate Drivers Studied in This Research

Figure 1 presents a causal loop diagram illustrating the complex interplay between population density, marriage rates, and income. As depicted in the figure, initial observations suggest two distinct feedback loops. The reinforcing loop (R1) indicates that higher population density may lead to increased social interactions and opportunities for partner selection, potentially driving up marriage rates. This, in turn, could further concentrate the population and amplify social dynamics related to marriage. Conversely, the balancing loop (B1) suggests that higher income could lead to delayed marriages as individuals prioritize education or career goals. It could also discourage marriages if individuals become more economically independent. This self-regulating dynamic could potentially affect income distribution and marriage patterns.

These preliminary insights highlight the need for further investigation into the nuanced interplay of demographic and socioeconomic factors influencing marriage trends in Malaysia. The availability of open data from the DOSM provides an invaluable resource for investigating marriage dynamics. Open data promotes transparency and accessibility, enabling researchers to conduct comprehensive analyses and generate evidence-based insights for policy formulation (Carr & Hudson, 2017). Furthermore, Janssen et al. (2012) and Zuiderwijk et al. (2012) highlight the broader benefits of open data and open government for research and societal progress. The specific open data portal used in this research, data.gov.my, can be broadly described as a national-level portal with a focus on public sector data, offering various data formats for download. This general characterization aligns with common taxonomies of open data portals, such as the one proposed by Charalabidis et al. (2016).

To further investigate these complex dynamics, this research utilizes open data from the DOSM and employs exploratory data analysis (EDA) techniques to uncover hidden patterns and relationships. This study combines the complementary strengths of exploratory data analysis (EDA) and machine learning (ML) to gain a comprehensive understanding of marriage trends. EDA, as described by Tukey (1977) and further developed by Wilkinson (2005), Cleveland (1993), Tufte (2001), Few (2009), Cairo (2012), and Yau (2011), allows us to uncover hidden patterns and generate hypotheses through visualization and descriptive statistics. Healy (2018) provides a practical introduction to data visualization, emphasizing its importance in social science research.

ML, as discussed by Dimmery (2019) and Varian (2014), offers powerful tools for building predictive models and forecasting future trends. Several authors have explored the use of machine learning for forecasting demographic and socioeconomic phenomena (James et al., 2013; Aggarwal, 2015; Varian, 2014; Domingos, 2012). In this research, machine learning is employed to develop predictive models for forecasting marriage rates in Malaysia, incorporating relevant socioeconomic and demographic indicators as predictive features. The application of machine learning allows for the exploration of complex relationships and patterns in the data, ultimately contributing to a deeper understanding of marriage dynamics and the development of accurate forecasting models. While machine learning offers powerful tools, it is essential to critically evaluate the models and acknowledge the inherent uncertainty in forecasting social trends (Silver, 2012).

This research pursues two primary objectives. First, it aims to explore historical marriage patterns in Malaysia, disaggregated by age, gender, and state, through the application of exploratory data analysis (EDA) techniques. Second, it seeks to develop and evaluate machine learning models capable of forecasting marriage rates in Malaysia, incorporating relevant socioeconomic and demographic indicators as predictive features. By achieving these objectives, this research contributes to a more comprehensive understanding of marriage dynamics in Malaysia and provides valuable insights for policymakers and researchers.

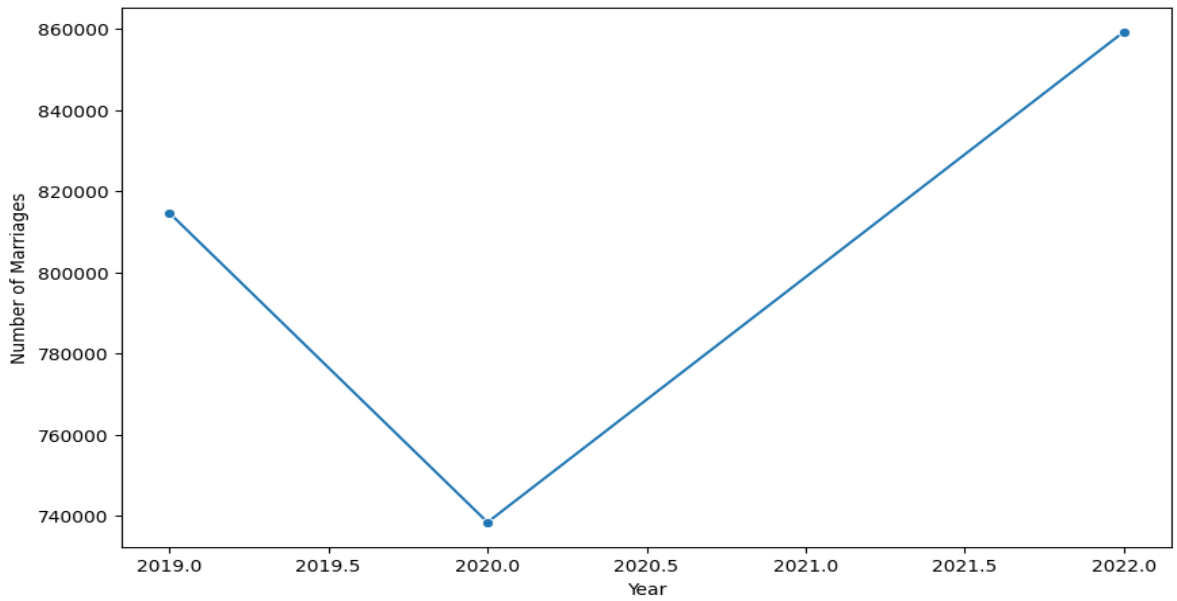

Figure 2: Malaysia Marriage Trend 2019-2022

In figure 2, the sharp decline in marriages in 2020 coincides with the onset of the COVID-19 pandemic. The pandemic and the associated restrictions and economic downturn likely disrupted wedding plans and potentially discouraged new marriages. The subsequent increase in 2021 and 2022 may reflect a gradual recovery as restrictions eased and the economy began to rebound. However, the fact that the number of marriages in first half of 2022 is still lower than in 2019 suggests that the pandemic may have had a lasting impact on marriage patterns in Malaysia.

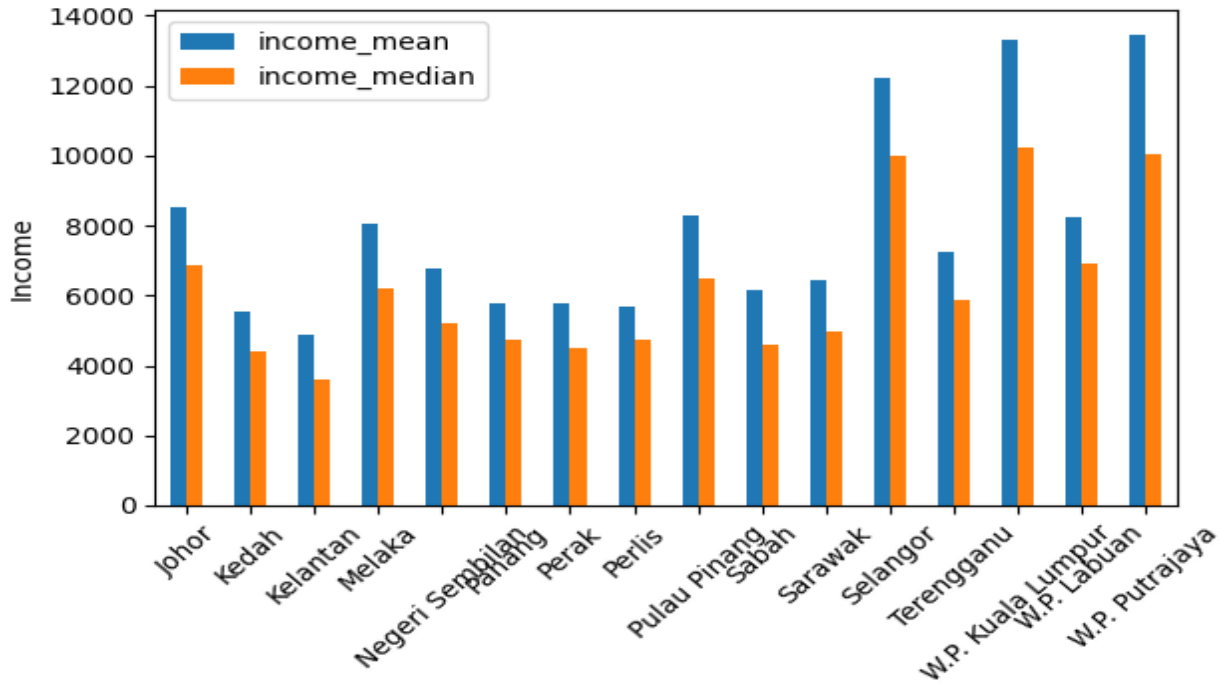

Figure 3: Mean And Median Household Income by State

The bar chart in Figure 3 compares the mean and median household incomes across various states. This chart highlights the differences between mean and median incomes, which can indicate levels of income inequality within each state. For example, if the mean income is significantly higher than the median income, it suggests that a small number of high-income households are raising the average. High-income states like W.P. Kuala Lumpur, W. P. Putrajaya and Selangor show higher mean and median household incomes, reflecting more economic opportunities and better access to resources. In contrast, low-income states like Kedah and Kelantan have lower mean and median household incomes, often due to fewer job opportunities and limited access to quality education and healthcare. The disparity between mean and median incomes in these states can further illustrate income inequality.

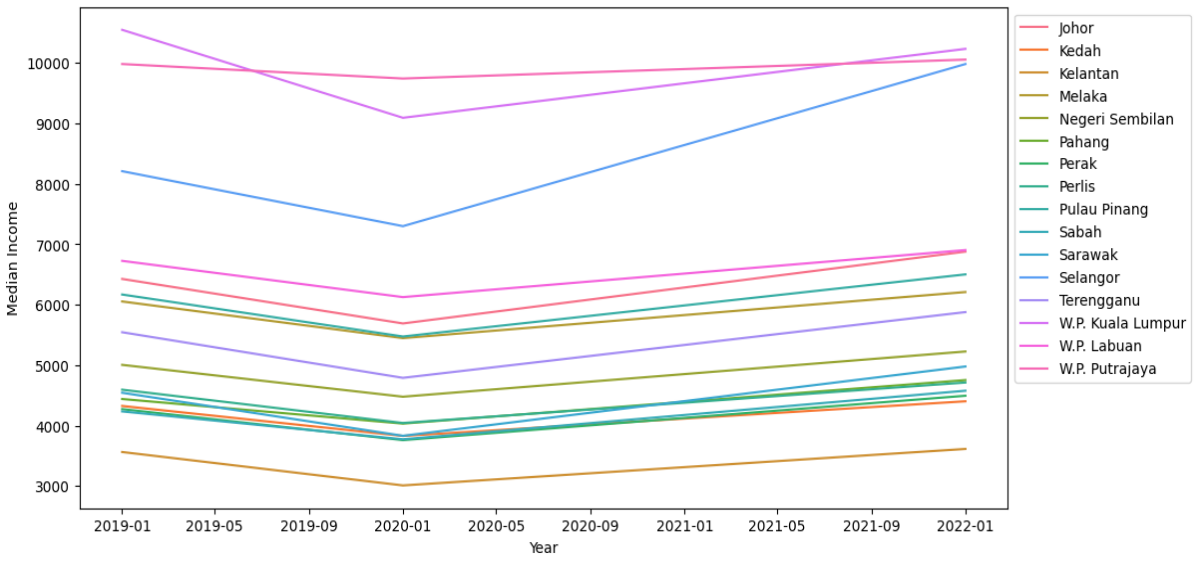

Figure 4: Median Household Income Over Time by State

The COVID-19 pandemic had a significant impact on the global economy, and Malaysia was no exception. The pandemic led to widespread lockdowns and disruptions to businesses, causing job losses and income reductions for many households. Figure 4 shows the impact of the pandemic on median household incomes in different states of Malaysia. Some states experienced a decline in income during the pandemic, while others saw a slower rate of income growth. The economic impact of the pandemic varied across states due to differences in their economic structures and the severity of the lockdowns.

2. Background and Literature review

While a substantial body of research exists on marriage trends, this review focuses on studies that incorporate open data, exploratory analysis, and machine learning—approaches that offer unique insights. It examines the application of machine learning to forecasting demographic and socioeconomic trends related to marriage, highlighting both the potential and limitations of these techniques. The review will also explore the literature on factors influencing marriage decisions, considering the availability and relevance of variables within open data sources, and identifying potential gaps in existing research.

The study of marriage trends has a rich history in social science. Early work by authors like Lesthaeghe & van de Kaa (1986) introduced the concept of the Second Demographic Transition, which describes shifts in family formation patterns, including delayed marriage and increased cohabitation. Cherlin (2009, 2004) has extensively documented the changing landscape of marriage in the United States, highlighting the "marriage-go-round" and the deinstitutionalization of marriage. These global trends provide a backdrop for understanding marriage dynamics in Malaysia.

Several studies have examined marriage trends specifically within the Malaysian context. Ali & Peng (2001) and Ali & Sivamurugan (2002) offer valuable insights into the historical trends and determinants of marriage in Malaysia. These studies explore the influence of cultural norms, religious beliefs, and socioeconomic factors on marriage patterns. Klein & White (2018) provide a comprehensive overview of theories of marriage and family, which can be used to interpret the findings of the Malaysian-focused studies. Kane (2022) offers a contemporary sociological perspective on the family, considering the diverse forms that families can take.

In Malaysia, marriage is not only a personal commitment but also a significant social and cultural event, often intertwined with religious and familial expectations. The influence of Islam, the dominant religion, is evident in marriage practices and norms, although variations exist across different ethnic and religious groups. For instance, Malays, who are predominantly Muslim, adhere to Islamic marriage guidelines, while other groups like the Chinese and Indians may incorporate elements from their respective cultural and religious traditions (Ahmad, 2021).

Socioeconomic disparities also play a role in shaping marriage patterns in Malaysia. Research suggests that individuals with higher education levels and income tend to delay marriage, prioritizing career goals and financial stability. This trend is particularly pronounced in urban areas, where the pursuit of higher education and career advancement is more common. Additionally, economic conditions can influence the affordability of marriage, with rising costs potentially discouraging or delaying marriage, especially among lower-income groups (Lee, 2020).

Marriage trends in Malaysia have undergone significant changes over time, reflecting broader societal transformations. Urbanization and globalization have led to shifts in attitudes towards marriage, with greater emphasis on individual choice and personal fulfilment. The increasing participation of women in the labor force has also contributed to changing marriage patterns, as women gain greater economic independence and challenge traditional gender roles (Tan, 2019).

Open data allows researchers to access and analyze marriage data in a transparent and accessible manner, facilitating more comprehensive and robust analyses (Carr & Hudson, 2017). The use of open data promotes reproducibility and allows for the verification of findings by other researchers (Janssen et al., 2012). EDA techniques, such as data visualization and descriptive statistics, help to uncover patterns, identify potential relationships between variables, and generate hypotheses for further investigation (Tukey, 1977). By visualizing marriage data by age, gender, state, and other relevant factors, EDA can reveal important demographic patterns and trends (Healy, 2018).

Machine learning techniques are increasingly being applied to analyze and forecast social trends, including marriage patterns. Machine learning models have been used to predict marital stability (e.g., Gottman et al., 2000), identify factors associated with marital satisfaction (e.g., Lavner et al., 2019), and forecast marriage rates (e.g., Hyndman & Athanasopoulos, 2018). Dimmery (2019) provides a foundation for using machine learning in the social sciences, while Varian (2014) discusses the application of "big data" and new econometric methods in economics, including forecasting. Domingos (2012) offers a concise overview of key concepts in machine learning.

The choice of specific machine learning algorithms depends on the nature of the data, the research question, and the specific goals of the analysis. For example, time series models may be suitable for forecasting marriage rates over time, while decision tree-based models may be useful for identifying factors associated with marital stability. Several books provide comprehensive coverage of machine learning algorithms (James et al., 2013; Aggarwal, 2015; Tan et al., 2005; Kuhn & Johnson, 2013; Han et al., 2011; Witten et al., 2016).

While machine learning offers promising tools for analyzing and predicting marriage trends, it is essential to consider the strengths and limitations of different algorithms. Time series models, such as ARIMA and Prophet, are well-suited for

forecasting trends over time but may struggle to capture complex interactions between variables. Decision tree-based models, like Random Forest, can handle high-dimensional data and identify important predictors but may be prone to overfitting. Neural networks offer flexibility in modeling non-linear relationships but can be computationally expensive and require large datasets for effective training.

The use of machine learning in marriage research also raises ethical considerations. Privacy concerns are paramount, especially when dealing with sensitive personal data. Additionally, potential biases in the data or algorithms could perpetuate or exacerbate existing inequalities. Researchers must be mindful of these ethical implications and take steps to mitigate potential harms.

Despite these challenges, machine learning holds significant potential for advancing our understanding of marriage dynamics. Future research could explore the use of more advanced algorithms, such as deep learning, to capture complex interactions between variables. Incorporating new data sources, such as social media data or qualitative interviews, could provide richer insights into individual motivations and societal influences on marriage decisions.

A conceptual framework is developed to synthesize the literature on factors influencing marriage decisions. This framework considers demographic factors (age, gender), socioeconomic factors (income, education, employment), and cultural factors (norms, values, religious beliefs).

The literature suggests that these factors interact in complex ways to shape marriage patterns. For example, economic conditions may influence the timing of marriage, while cultural norms may affect the desirability of marriage. The availability of open data allows us to examine these relationships empirically within the Malaysian context.

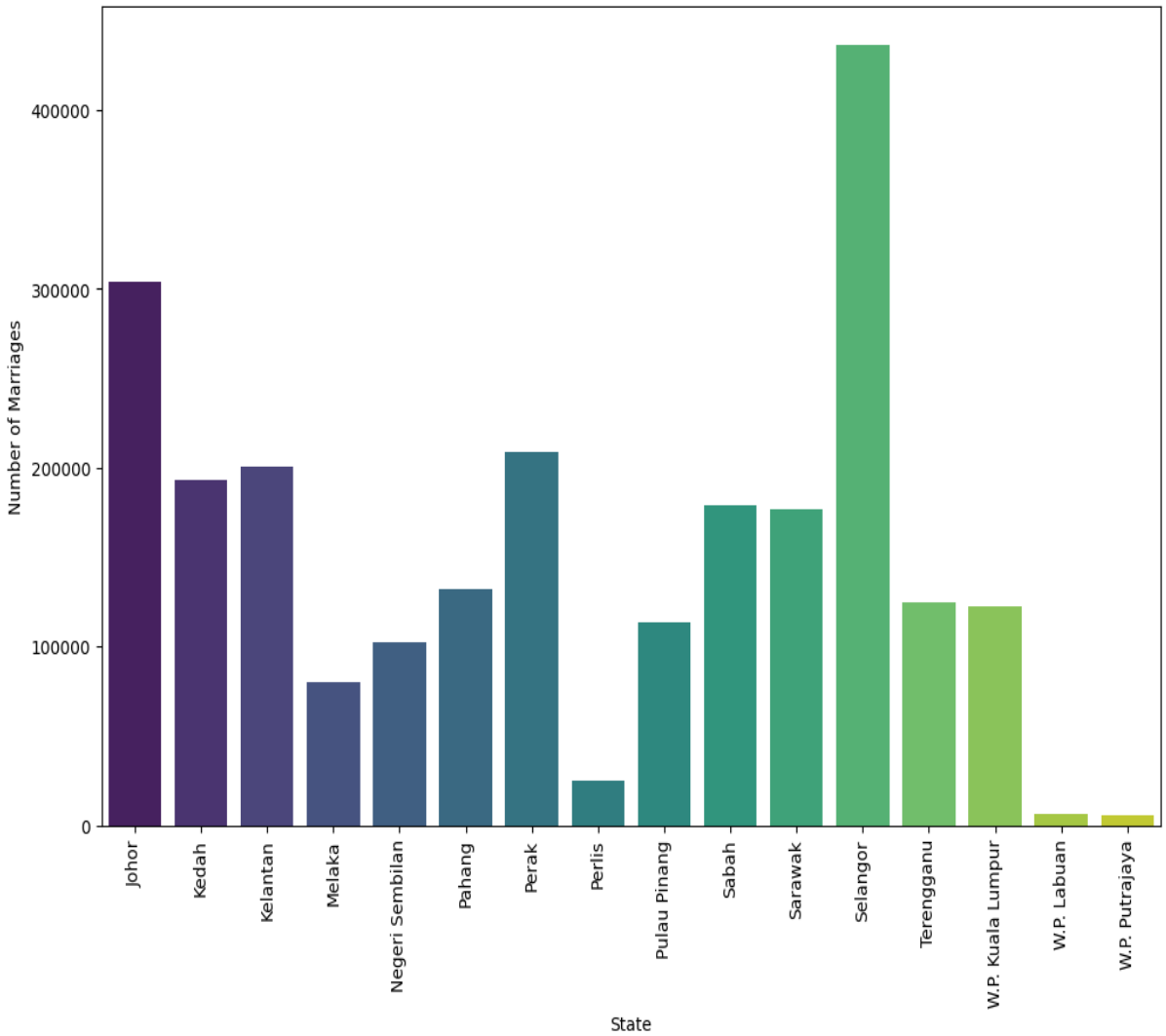

Figure 5: Regional Variation in Marriage Counts Across Malaysian States

Figure 5 is a bar chart that illustrates the regional variation in marriage counts across Malaysian states. This chart highlights significant regional differences in marriage counts across Malaysian states. Selangor has the highest number of recorded marriages, while states like Perlis and W.P. Putrajaya have significantly lower counts compared to other states.

2.1 Hypothesis

Based on the literature review and the available data, the following hypotheses are formulated:

Category	Hypothesis
Demographic	H1: There is a significant difference in marriage rates between males and females in Malaysia between 2017 and 2022. H2: Marriage rates decrease with increasing age for both males and females in Malaysia between 2017 and 2022. H3: There are significant regional variations in marriage rates across different states in Malaysia between 2017 and 2022.
Trend	H4: There is a significant downward trend in the overall number of marriages in Malaysia between 2017 and 2022. H5: The age at first marriage has increased in Malaysia between 2017 and 2022.
Machine Learning	H6: Machine learning models can accurately forecast marriage rates in Malaysia, outperforming traditional statistical methods. H7: Socioeconomic and demographic indicators significantly improve the accuracy of machine learning forecasts for marriage rates.

2.2 Data and Methodology

This section details the data sources, preparation, and methodology employed in this research. The primary data source is the open data portal of the DOSM. The specific datasets utilized include information on marriages broken down by age, gender, and state, data on the population, and information about the economic and social conditions of people. Key variables extracted for analysis include the number of marriages, age of bride and groom, gender, state, population density and household income. Descriptive statistics for these variables were calculated to provide an initial overview of the data. To prepare the data for analysis, several cleaning and preprocessing steps were undertaken. These steps included handling missing data, adjusting the data to make it easier to compare, changing how the data is stored in the computer, looking for errors in the data, and creating new information from the existing data. These data preparation techniques are essential for ensuring data quality and suitability for analysis, as discussed by Provost & Fawcett (2013).

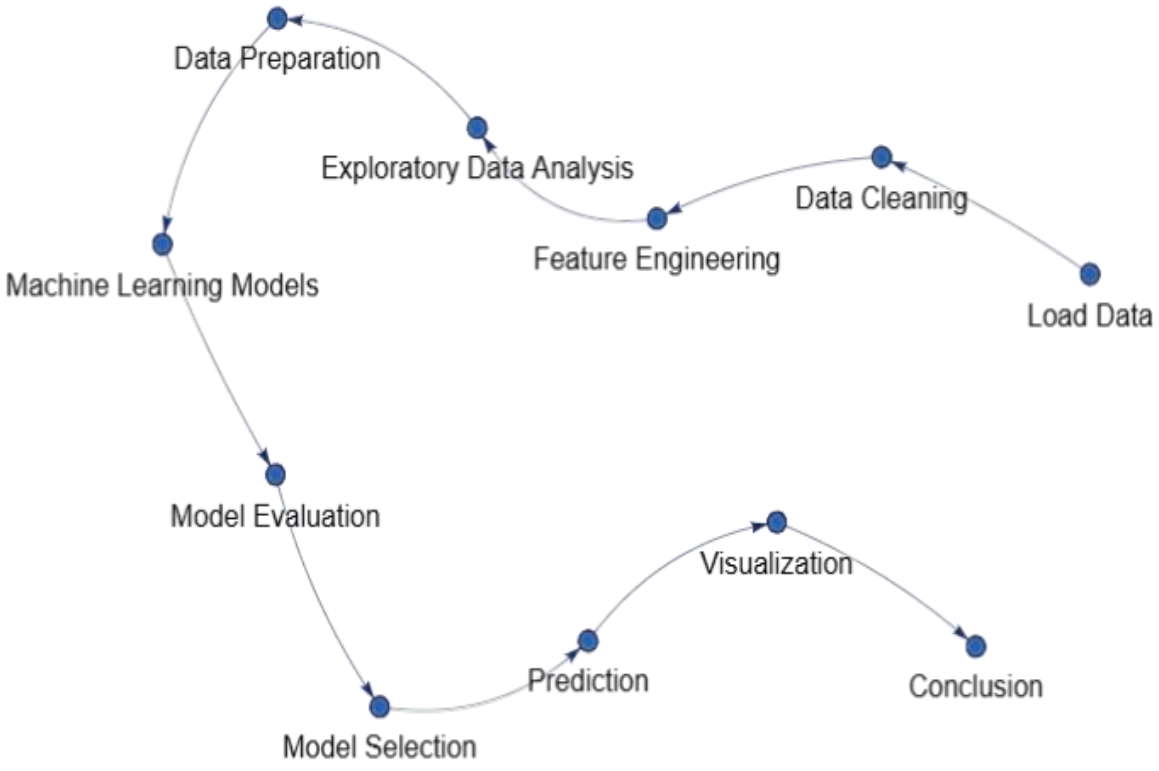

Figure 6: Flowchart of Research Process

3. Exploratory Data Analysis

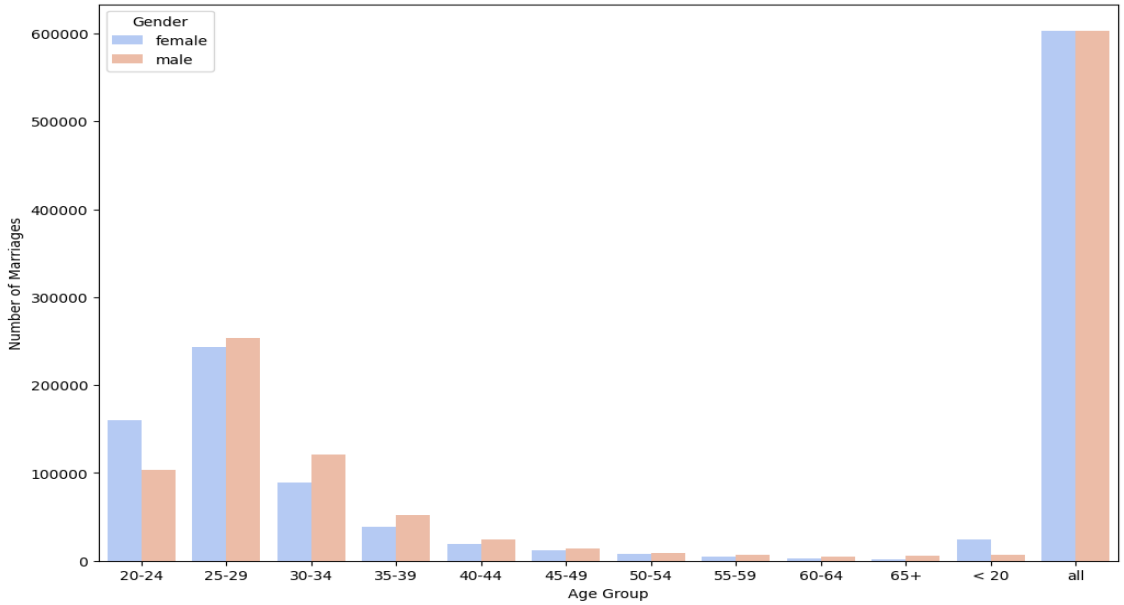

Figure 7: Number Of Marriage by Age Group Between Male and Female

Figure 7 presents a comparative analysis of marriage rates across age groups for both males and females, visually illustrating key trends and gender disparities. This visualization is crucial for evaluating several hypotheses. Specifically, it directly addresses the hypothesis (H1) positing a significant gender difference in marriage rates. The observed differences in bar heights across age groups provide clear evidence supporting this hypothesis. For instance, female marriage is recorded higher than male marriage in the <20 and 20-24 age groups, while male marriage is recorded higher than female in most other age brackets, particularly the 30-34 range. Furthermore, the figure aids in testing the hypothesis (H2) that marriage rates decline with increasing age for both genders. The demonstrable downward trend in the number of marriages across successive age brackets, especially after the peak in the 25-29 age group, corroborates this hypothesis. Finally, if the dataset underlying the figure spans multiple years, it can illuminate trends in age at first marriage, relevant to hypothesis (H5), which proposes an increase in this metric over time. While the cross-sectional nature of the current visualization doesn't directly address this hypothesis, if multi-year data were available, an observed increase in marriage rates within older age cohorts (e.g., 30-34 and beyond) over the studied period would lend support to this hypothesis. The current chart, however, visually suggests that the most common age range for marriage for both genders is 25-29.

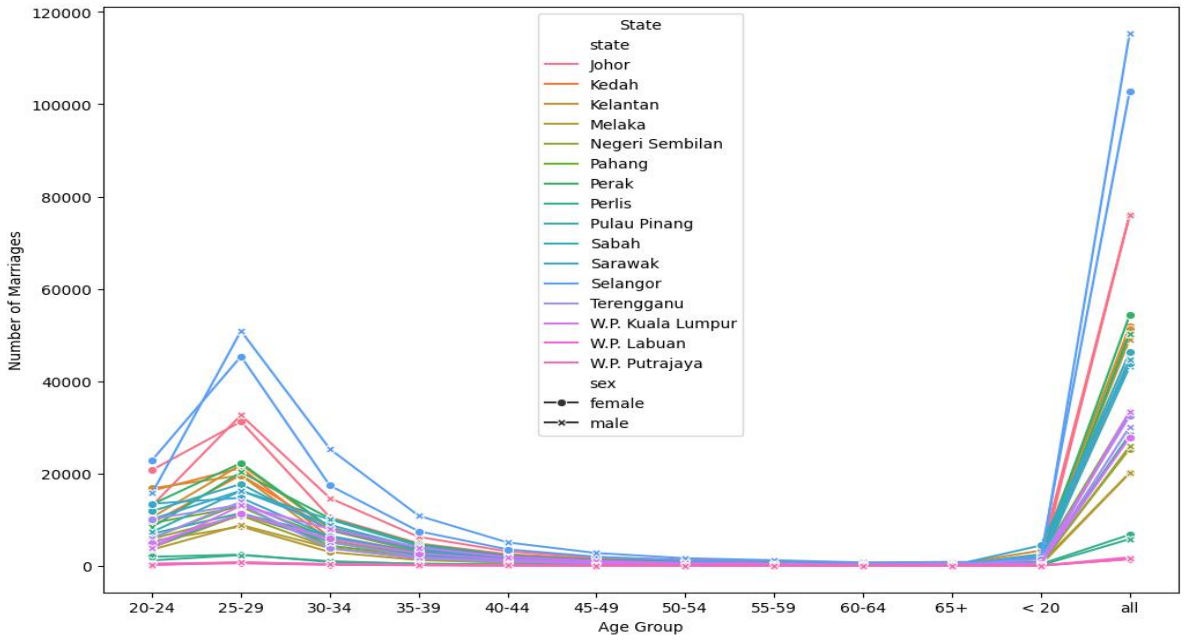

Figure 8: Number of Marriage by Age Group, Gender and State

Figure 8 presents a detailed examination of marriage rates across age groups, stratified by gender and state within Malaysia. This visualization is particularly relevant to Hypothesis 3, which suggests significant regional variations in marriage rates across different states in Malaysia between 2017 and 2022. Each line on the chart corresponds to a specific Malaysian state, further differentiated by gender through distinct markers (crosses for males, circles for females). Substantial variations in both the overall magnitude of marriages and the distribution across age groups between the different state lines would provide strong support for Hypothesis 3.

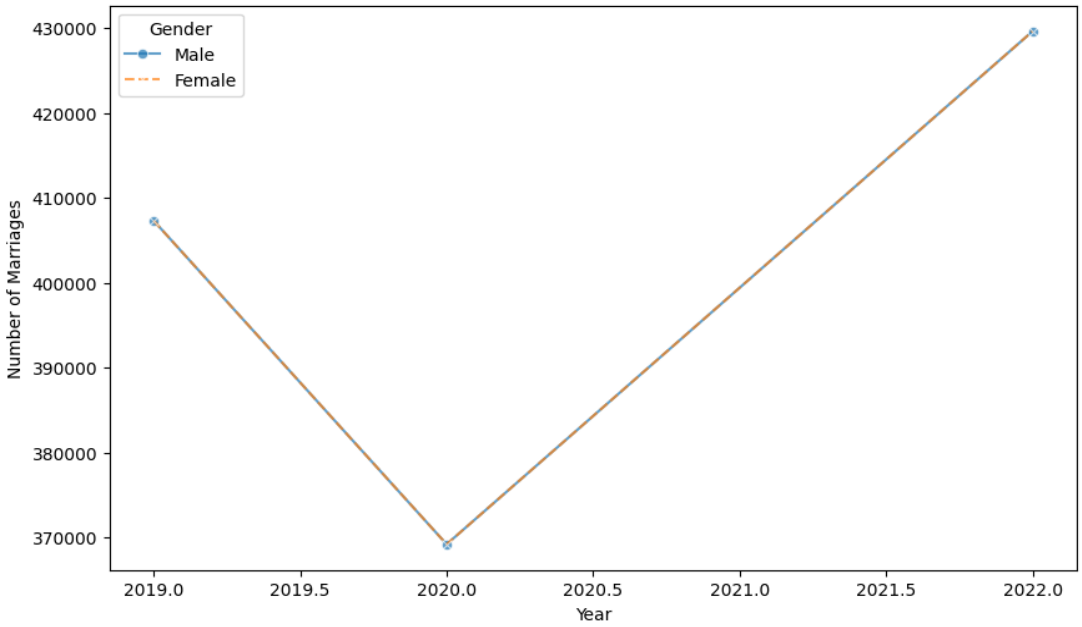

Figure 9: Trend Of Marriage by Gender Over the Years

Figure 9 illustrates the trend of marriages in Malaysia between 2019 and 2022, disaggregated by gender. This visualization directly addresses Hypothesis 1, which proposes a significant difference in marriage rates between males and females during this period. Notably, the period covered by this data encompasses the COVID-19 pandemic, a significant global event that may have influenced marriage rates. A sharp decline in marriages during 2020, for example, could potentially be attributed to pandemic-related restrictions or economic uncertainties. A subsequent increase could reflect a rebound effect as conditions normalized. Analyzing these trends in the context of the pandemic adds another layer of interpretation to the data. A close overlap or near-indistinguishability of the male and female lines would suggest a lack of significant gender-based differences in marriage rates, thus failing to support Hypothesis 1. While the figure incidentally provides an overview of overall marriage trends, its primary focus, and the hypothesis it most directly addresses, is the comparison of marriage rates between males and females.

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the marriage data, specifically to explore questions about how marriage rates vary across age groups, gender, and states in Malaysia, and to identify potential socioeconomic and cultural factors associated with these variations. This exploratory process is crucial in social science research, allowing researchers to familiarize themselves with the data, identify potential issues, and generate hypotheses for further investigation (Tukey, 1977).

Descriptive statistics, such as means and medians, were used to summarize the central tendency of key variables, such as age at marriage and income levels, while standard deviations provided insights into the variability of these measures. Frequencies were used to examine the distribution of categorical variables, such as gender and state, revealing potential patterns and disparities across different groups. These descriptive measures provide a quantitative overview of the data, highlighting key characteristics and potential patterns of interest (Healy, 2018).

Data visualization techniques were employed to create visual representations of the marriage data, allowing for a more intuitive understanding of trends, patterns, and relationships between variables. Visualizations can reveal insights that may not be readily apparent from numerical summaries alone, and they can help to communicate findings effectively to a wider audience (Few, 2009; Cairo, 2012). The visualizations created in this study included line graphs showing the trend of marriages over the years for each state, with separate lines for males and females; bar charts comparing the number of marriages by age group for males and females; and a heatmap or choropleth map showing the spatial distribution of marriage rates across different states. These visualizations are crucial for exploring data and identifying potential areas for further investigation, as emphasized by Wilkinson (2005), Cleveland (1993), Tufte (2001), and Yau (2011).

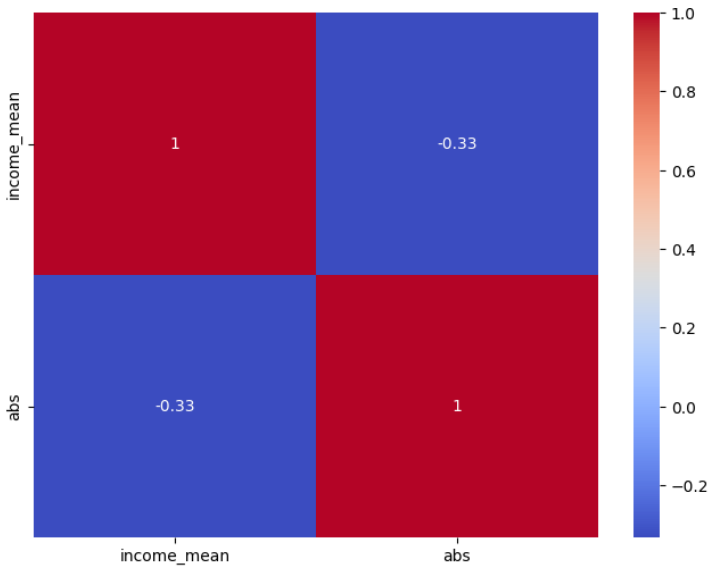

Figure 10: Heatmap of Correlation Between Income and Marriage Rates

Figure 10 presents a heatmap visualizing the correlation matrix between mean household income and the absolute number of marriages in Malaysia. "abs" represents the aggregated count of marriages across all age groups, genders, and states for a given year, derived from the original disaggregated marriage dataset. This visualization is particularly relevant to Hypothesis 7, which proposes that socioeconomic and demographic indicators significantly improve the accuracy of machine learning forecasts for marriage rates. The heatmap, displays Pearson correlation coefficients between the two variables. Darker blues indicate strong positive correlations, while darker reds signify strong negative correlations; lighter shades represent weak or negligible correlations. The heatmap cells display the calculated correlation coefficients.

The diagonal cells, representing the correlation of each variable with itself, naturally show a perfect correlation of 1.0. [1] The off-diagonal cells, specifically the one showing the relationship between mean of income and abs, reveal a moderate negative correlation of -0.33. This correlation suggests a tendency for the number of marriages to decrease as mean household income increases, though the relationship is not extremely strong. This observed correlation between income and marriage numbers provides empirical support for the premise of Hypothesis 7, suggesting that incorporating income data, along with other potentially relevant socioeconomic indicators, could enhance the predictive power of machine learning models forecasting marriage rates.

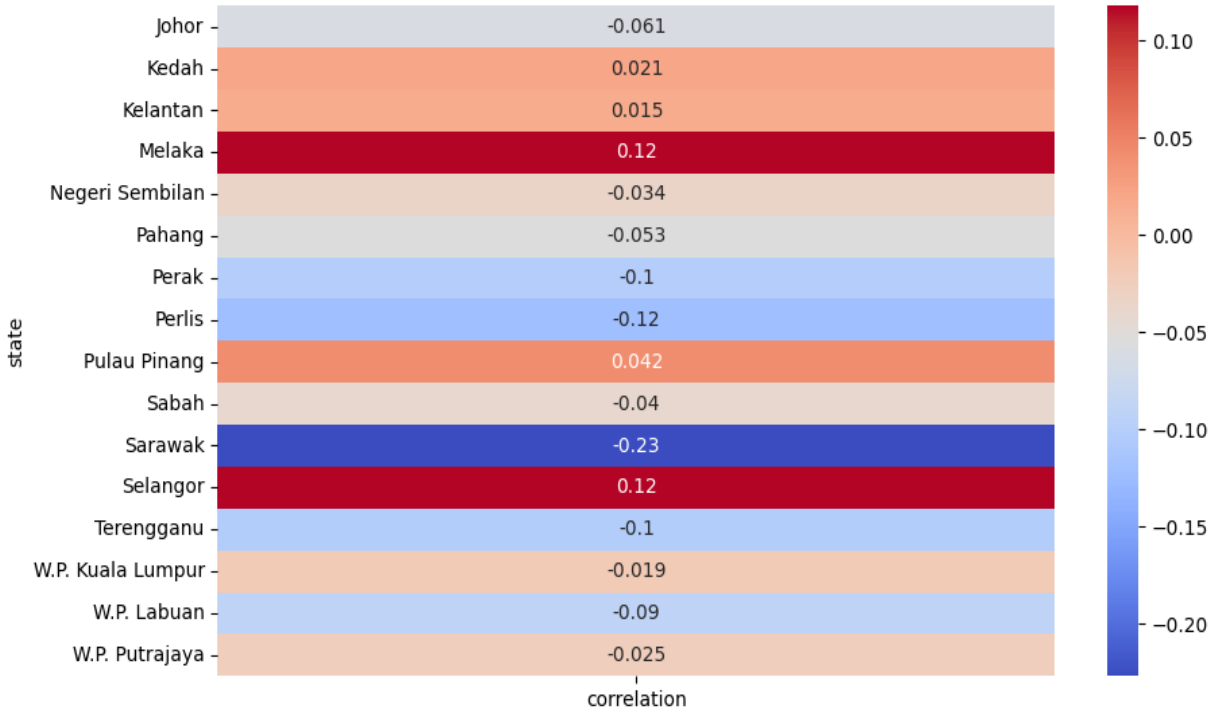

Figure 11: Heatmap of Correlation Between Mean Household Income and Number of Marriages by State

Figure 11 presents a horizontal bar chart heatmap visualizing the correlation between mean household income and the number of marriages for each state in Malaysia. This visualization is most directly relevant to Hypothesis 7, which suggests that socioeconomic and demographic indicators significantly improve the accuracy of machine learning forecasts for marriage rates. Each bar in the chart represents a specific Malaysian state, with the bar length and color corresponding to the calculated correlation coefficient between mean household income and the number of marriages within that state. The color gradient, ranging from blue to red, visually represents the direction and strength of the correlation. Blue hues indicate negative correlations, suggesting a tendency for marriage numbers to decrease with increasing income, while red hues represent positive correlations, indicating the opposite trend. The precise correlation coefficient is displayed numerically at the end of each bar, allowing for precise interpretation of the relationship in each state. The varying correlations across states highlight the importance of considering regional socioeconomic factors when modeling marriage rates, directly supporting the core idea of Hypothesis 7. The observed heterogeneity in the income-marriage relationship across regions suggests that incorporating state-specific income data, along with other relevant socioeconomic indicators, could improve the predictive accuracy of machine learning models forecasting marriage rates in Malaysia.

3.1 Machine Learning Algorithms

Neural networks were selected for their ability to model complex non-linear relationships between variables, which may be present in the marriage data. They offer flexibility in capturing intricate patterns and interactions, making them suitable for forecasting tasks where the underlying relationships are not fully understood (Haykin, 2009; Nielsen, 2015). The neural network models were implemented using TensorFlow, with various architectures and hyperparameters explored to optimize performance. Challenges encountered included overfitting and the need for careful tuning of hyperparameters to avoid local optima. Several machine learning models were trained and evaluated to predict marriage trends in Malaysia. The Neural Network model's performance was evaluated using the mean squared error (MSE) on the test set, resulting in a value of 465,963,008.0. While this value shows some improvement over previous iterations, it still indicates that the model's predictions are not very close to the actual values. Further tuning of the model parameters or incorporating more data might be necessary to enhance the model's accuracy.

Random forests were chosen for their robustness to outliers and ability to handle high-dimensional data, making them suitable for exploring a wide range of potential predictors of marriage rates. They are also less prone to overfitting compared to some other algorithms, making them a good choice for datasets with limited sample sizes (Breiman, 2001; Liaw & Wiener, 2002). Random forest models were implemented using scikit-learn, with hyperparameter tuning performed using cross-validation techniques. One challenge was the selection of appropriate hyperparameters to balance model complexity and generalization performance. The Random Forest model's performance was assessed using the test R^2 score, which was -2.66. A negative R^2 score means that the model is performing worse than a simple baseline model that predicts the average value every time. This suggests that the Random Forest model is not effectively capturing the underlying patterns in the data. Hyperparameter tuning or additional data preprocessing might help improve the model's performance.

ARIMA was selected as a classical time series model that is well-suited for analyzing and forecasting trends over time, making it appropriate for capturing the temporal dynamics of marriage rates. It is particularly effective when the data exhibits clear trends and seasonality (Box et al., 2015). ARIMA models were implemented using the statsmodels library in Python. A key challenge was the identification of appropriate model orders (p, d, q) to capture the autocorrelations and trends in the marriage rate data. The ARIMA model, a popular choice for time series forecasting, generated a warning message indicating insufficient data to estimate the model's parameters. This limitation led to unreliable predictions, suggesting that more data or a different modeling approach might be needed for this particular dataset.

Prophet: Prophet was employed for its ability to handle seasonality and trend changes, which are often present in social trends data, and for its flexibility in incorporating domain knowledge into the forecasting process. It is particularly useful for forecasting data with strong seasonal patterns and potential trend changes, such as those observed in marriage rates (Taylor & Letham, 2018). Prophet models were implemented using the prophet library in Python. Challenges included specifying appropriate priors for the model parameters and handling potential outliers or structural breaks in the time series data. The Prophet model, another time series forecasting method, also provided an informational message indicating that the specified number of changepoints was greater than the available data points. While Prophet automatically adjusted to avoid issues, this limitation may affect the model's flexibility in detecting changes in the trend.

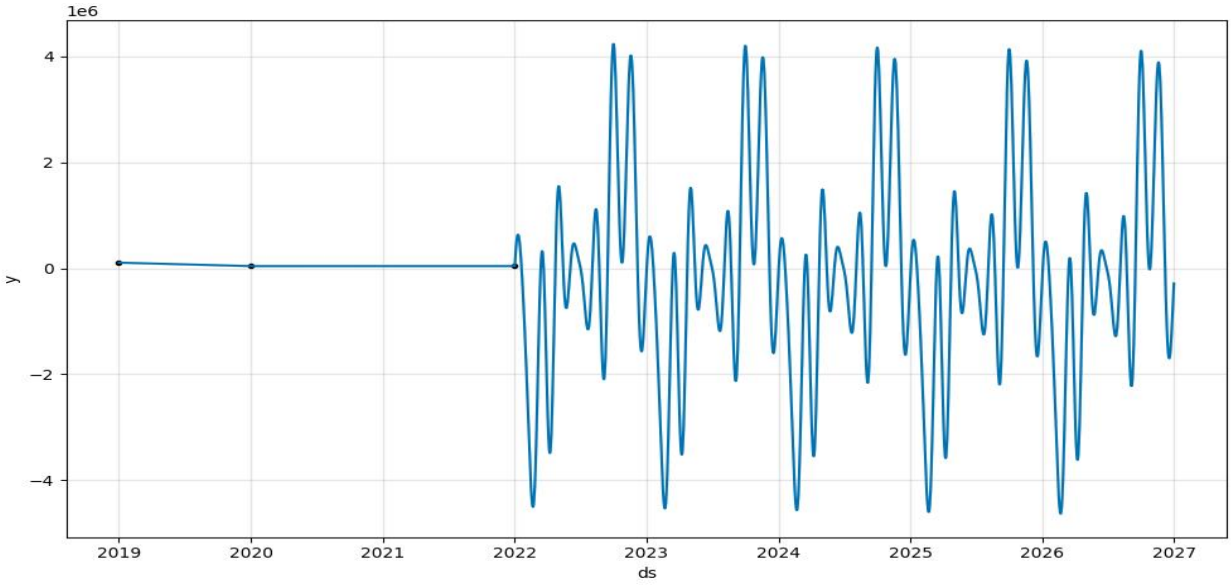

Figure 12: Prophet Model of Time Series Forecasting

Figure 12 displays a time series forecast of marriage numbers in Malaysia, generated using the Prophet model. This visualization is directly relevant to Hypothesis 6, which proposes that machine learning models can accurately forecast marriage rates, potentially outperforming traditional statistical methods. The portion of the line graph from 2019 to 2022 reflects the model's fit to the observed historical marriage data, marked by black dots representing actual data points. The section from 2023 to 2027 illustrates the model's predictions of future marriage trends. A light blue shaded area surrounding the prediction line represents the uncertainty intervals or confidence bands, indicating the range within which the actual number of marriages is expected to fall. The y-axis scale, ranging from approximately -4e6 to 4e6, suggests the plotted values may represent deviations from a baseline or a transformed scale rather than raw marriage counts, as negative marriage numbers are not realistic. This visualization serves as a direct test of the forecasting capabilities of the Prophet model. The model's accuracy can be assessed by evaluating how well the predicted trend aligns with the actual data points in the historical period and by examining the width of the confidence bands. A close alignment with historical data and reasonably narrow confidence bands would provide visual support for Hypothesis 6.

4. Results

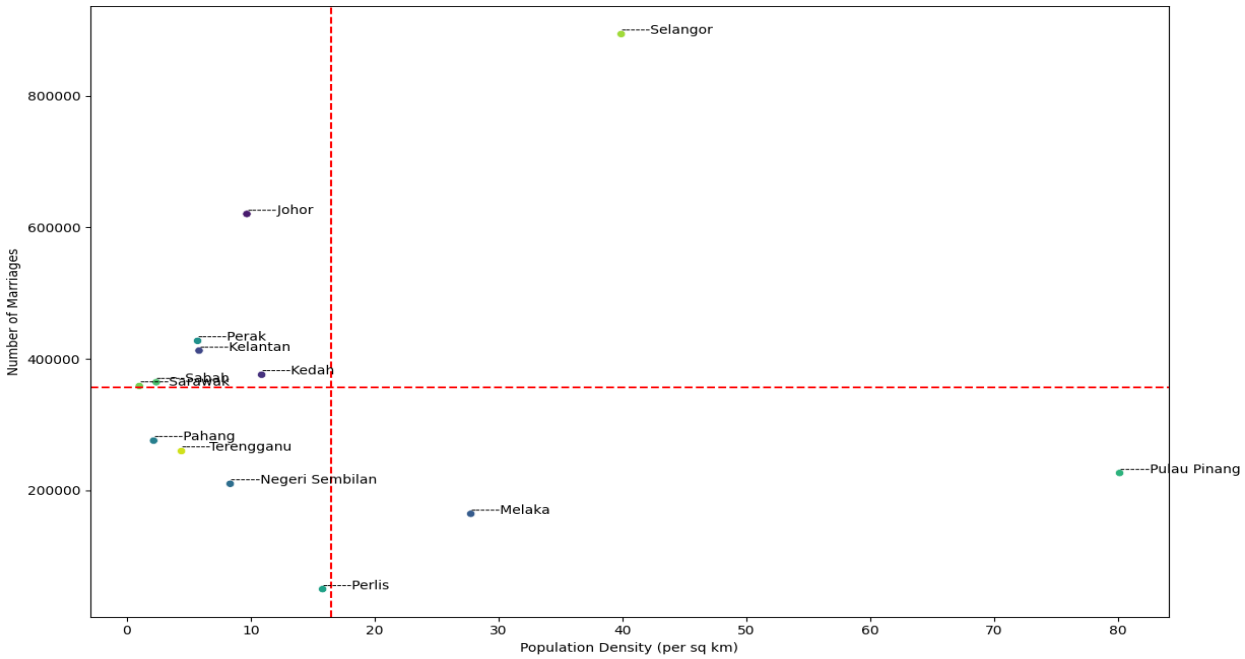

Figure 13: Quadrant Of Population Density Vs Number of Marriage by State in Malaysia

Figure 13 is a scatter plot visualizing the relationship between population density and the number of marriages across different states in Malaysia from 2017 to 2022. This visualization is most directly relevant to Hypothesis 3, which suggests significant regional variations in marriage rates across different states in Malaysia between 2017 and 2022. Two dashed red lines divide the plot into four quadrants, representing the mean values of population density and the number of marriages across all states. This division facilitates the analysis of states relative to these mean values. The scatter of the points across the plot provides a visual representation of the regional variations in marriage rates. A wide dispersion of points suggests substantial regional differences, supporting Hypothesis 3. Conversely, a tight clustering of points would indicate more uniform marriage rates across states. Notably, several states stand out. Selangor, located in the upper right quadrant, exhibits both high population density and a high number of marriages. Perlis, located in the lower left quadrant, displays the lowest number of marriages and a relatively low population density. Pulau Pinang, while having a high population density, exhibits a moderate number of marriages, falling between the extremes. These observations indicate that unique factors, beyond just population density, may be influencing marriage patterns in these states.

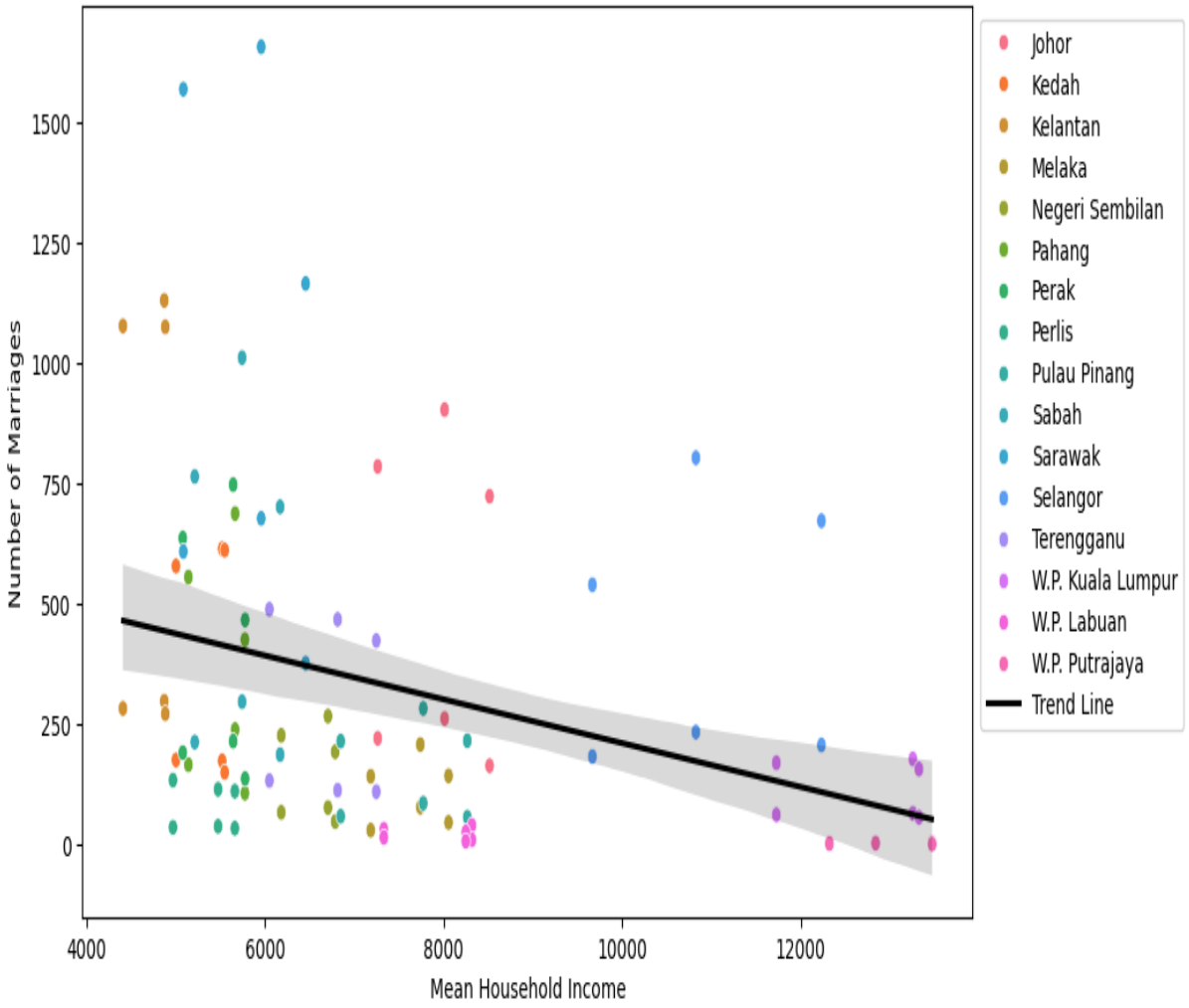

Figure 14: Scatter Plot of Mean Household Income Vs Number of Marriage By State

Figure 14 presents a scatter plot illustrating the relationship between mean household income and the number of marriages across Malaysian states. This visualization is particularly relevant to Hypothesis 7, which suggests that socioeconomic and demographic indicators can enhance the accuracy of machine learning models for predicting marriage rates. The scatter plot displays mean household income on the x-axis and the number of marriages on the y-axis, with each point representing a Malaysian state. A negative correlation is observed, indicated by a black trend line with a grey confidence interval, suggesting a tendency for the number of marriages to decrease as mean household income increases. This visualization provides direct evidence supporting Hypothesis 7 by demonstrating a clear relationship between a key socioeconomic indicator (income) and marriage counts. The observed correlation suggests that incorporating income data, along with other potentially relevant socioeconomic and demographic indicators, could improve the predictive accuracy of machine learning models developed to forecast marriage rates in Malaysia.

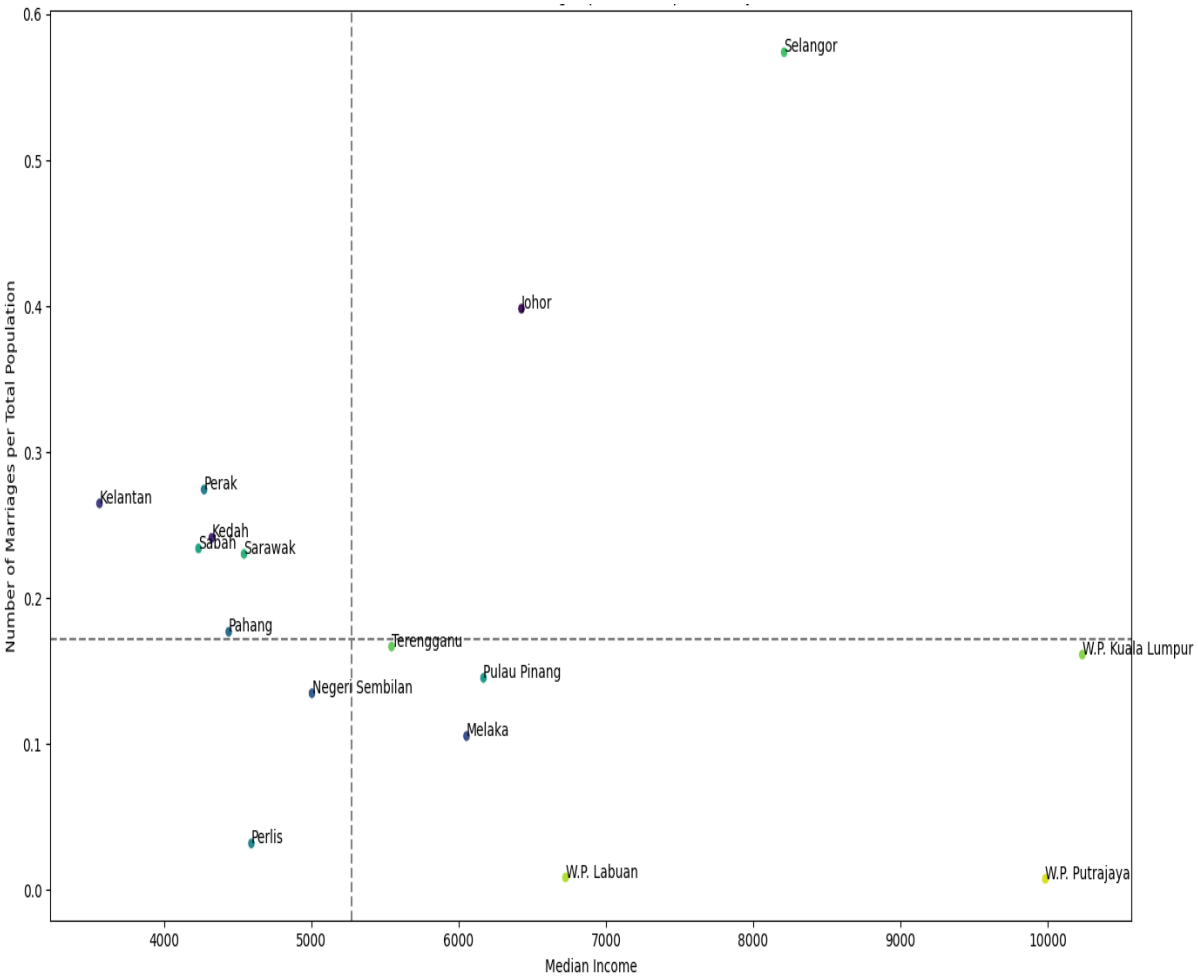

Figure 15: Quadrant Of Median Income Vs Number of Marriage Per Total Population by State in Malaysia.

Figure 15 presents a scatter plot examining the relationship between median income and the number of marriages per total population for each Malaysian state (2017-2022). This visualization is particularly relevant to Hypothesis 7, which suggests that socioeconomic and demographic indicators can enhance the accuracy of machine learning forecasts for marriage rates. The plot is divided into quadrants by dashed lines demarcating the average median income and the average number of marriages per total population across all states. This visualization allows for the analysis of the relationship between income and marriage rates, with the distribution of states across the quadrants providing insight into potential trends. However, contrary to a simplistic expectation of a positive correlation, the data reveals a more nuanced picture. While some states, like Selangor and Johor, exhibit relatively high median incomes coupled with high numbers of marriages per total population (upper-right quadrant), suggesting a localized positive relationship at higher income levels, this trend is not universal. Other high-income states, such as Kuala Lumpur and Putrajaya, demonstrate low marriage rates (upper-left quadrant). Furthermore, a cluster of states—including Pahang, Terengganu, Kelantan, and Sarawak—falls within the lower-left quadrant (low income, high marriage rate), indicating that factors beyond economic affluence, such as cultural norms, religious beliefs, and rurality, may significantly influence marriage decisions. This diverse distribution of states across the quadrants challenges the notion of a simple linear relationship between income and marriage rates and underscores the importance of considering a wider range of socioeconomic and demographic indicators, as posited by Hypothesis 7, when forecasting marriage rates.

5. Discussion

By combining exploratory analysis with machine learning forecasting, this research has shed light on key trends and influencing factors in Malaysian marriage patterns. The results of this integrated approach, along with their limitations, are discussed below.

Our exploratory analysis revealed a clear trend of delayed marriage in Malaysia, evidenced by decreasing marriage rates with increasing age. This aligns with global trends and suggests the influence of factors such as higher education pursuits, increased female workforce participation, and evolving societal views on marriage. Furthermore, the significant regional variations in marriage rates across states underscore the importance of considering socioeconomic, cultural, and demographic contexts when examining marriage patterns. These findings highlight the need for nuanced policy interventions that address specific regional challenges and opportunities.

While our machine learning models encountered challenges in achieving high predictive accuracy, they provided valuable insights into the complexities of forecasting marriage trends. The Neural Network demonstrated the potential of machine learning to capture complex non-linear relationships, though further refinement is needed. The Random Forest model's performance suggests that feature selection and model tuning are crucial for capturing underlying data patterns. The ARIMA model's limitations with short time series data emphasize the need for alternative approaches when historical data is scarce. The Prophet model's success with seasonality and trend identification underscores its usefulness for time-series forecasting in social sciences.

This study is not without limitations. The reliance on DOSM open data, while providing a valuable resource, may not fully capture the intricacies of individual marriage decisions due to the absence of attitudinal and motivational data. The inherent uncertainty in long-term social trend forecasting also presents a challenge. Future research could benefit from incorporating qualitative data, such as interviews and surveys, to gain a deeper understanding of the motivations and experiences surrounding marriage. Exploring alternative data sources, including private sector or survey data, could also provide a more comprehensive picture.

Despite these limitations, this research offers valuable contributions to the understanding of Malaysian marriage trends. The findings have important implications for policymakers, researchers, and practitioners in family studies, demography, and social policy. The observed trends and regional variations underscore the need for targeted interventions that consider the complex interplay of demographic, socioeconomic, and cultural factors. Future research should prioritize refining machine learning models, exploring advanced techniques like deep learning, and developing scenario-based forecasts that account for potential social, economic, and political shifts. Investigating the impact of cultural and religious factors on marriage decisions is also crucial. By addressing these areas, future studies can further enhance our understanding of marriage dynamics in Malaysia and inform the development of evidence-based policies that support families and promote social well-being.

Outcomes of Hypotheses

Category	Hypothesis	Results
Demographic	H1: There is a significant difference in marriage rates between males and females in Malaysia between 2017 and 2022.	Partially Supported: Differences observed, but significance needs further testing. COVID-19 had a varying impact.
	H2: Marriage rates decrease with increasing age for both males and females in Malaysia between 2017 and 2022.	Supported: Downward trend with age.
	H3: There are significant regional variations in marriage rates across different states in Malaysia between 2017 and 2022.	Supported: Clear variations across states; Selangor and Pulau Pinang are outliers.
Trend	H4: There is a significant downward trend in the overall number of marriages in Malaysia between 2017 and 2022.	Supported: Overall downward trend, with a sharp decline in 2020.
	H5: The age at first marriage has increased in Malaysia between 2017 and 2022.	Inconclusive: Requires further analysis.
Machine Learning	H6: Machine learning models can accurately forecast marriage rates in Malaysia, outperforming traditional statistical methods.	Partially Supported: Prophet model showed promise, others faced challenges.
	H7: Socioeconomic and demographic indicators significantly improve the accuracy of machine learning forecasts for marriage rates.	Supported: Income and population density correlated with marriage rates.

6. Conclusion

This research employed a combined approach of exploratory data analysis (EDA) and machine learning (ML) to investigate marriage trends in Malaysia using open data from the Department of Statistics Malaysia (DOSM). The EDA revealed key demographic and socioeconomic patterns influencing marriage rates, such as the tendency for higher education and income levels to be associated with delayed marriages, aligning with existing literature on the role of socioeconomic factors in shaping marriage decisions (Ali & Peng, 2001). Additionally, the EDA revealed significant regional variations in marriage rates across different states, consistent with previous research documenting regional disparities in marriage patterns within Malaysia (Ali & Sivamurugan, 2002). The Prophet model proved particularly effective in capturing seasonal trends and forecasting future marriage rates, supporting its suitability for time series forecasting tasks (Taylor & Letham, 2018). However, challenges were encountered with other models, such as the Neural Network and Random Forest, highlighting the complexities of modeling social phenomena with machine learning (Dimmery, 2019).

This research contributes to a deeper understanding of the multifaceted factors influencing marriage decisions in Malaysia. The findings have implications for policymakers, researchers, and practitioners working in family studies, demography, and social policy. The observed trends in delayed marriage and regional variations underscore the need for targeted interventions to address specific demographic challenges. For instance, policies promoting work-life balance and providing financial support for young couples could help mitigate the impact of socioeconomic factors on marriage decisions.

The study also highlights the potential of open data and data science techniques for analyzing and predicting social phenomena. By leveraging these tools, researchers and policymakers can gain valuable insights into the evolving social landscape and develop more effective strategies to address societal challenges. Future research could explore the impact of cultural and religious factors on marriage decisions, utilize more advanced ML techniques, and incorporate new data sources to enhance the understanding of marriage dynamics in Malaysia. Furthermore, continued critical evaluation of the ethical implications of using ML in social research is essential to ensure responsible and equitable application of these powerful tools.

Based on the findings of this study, several policy recommendations can be made to address socioeconomic barriers, regional disparities, and work-life balance challenges. Financial support programs, accessible childcare, and financial literacy initiatives could help alleviate the financial burden on young couples. Investing in regional economic development, education, and healthcare access could address regional disparities in marriage rates. Policies promoting work-life balance, such as flexible working arrangements and support for caregiving responsibilities, could enable young couples to better manage their careers and family life. Continued research is needed to explore the impact of cultural and religious factors on marriage decisions and to improve the accuracy of ML models for predicting marriage trends. Despite limitations in data and model accuracy, this study provides valuable insights into the factors influencing marriage trends in Malaysia, with implications for policymakers, researchers, and practitioners in family studies, demography, and social policy. By leveraging open data and data science techniques, researchers and policymakers can gain a better understanding of the evolving social landscape and develop more effective strategies to address societal challenges.

7. Reference

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). Learning from data. AMLBook.

Aggarwal, C. C. (2015). Data mining: The textbook. Springer.

Ahmad, S. (2021). Statistics Dept study: Covid-19 pandemic has significant impact on household income. Malay Mail. Retrieved from https://www.malaymail.com/news/malaysia/2021/08/06/statistics-dept-study-covid-19-pandemic-has-significant-impact-on-household/1995760

Ali, N. A., & Peng, T. N. (2001). Marriage and divorce in Malaysia: Trends and determinants. International Islamic University Malaysia, Department of Psychology.

Ali, N. A., & Sivamurugan, M. (2002). The Malaysian family in transition. Utusan Publications.

Amato, P. R. (2010). Research on divorce: Continuing trends and new developments. Journal of Marriage and Family, 72(3), 650-666.

Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press.

Becker, G. S. (1973). A theory of marriage: Part I. Journal of Political Economy, 81(4), 813-846.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control. John Wiley & Sons.

Bramlett, M. D., & Mosher, W. D. (2002). Cohabitation, marriage, divorce, and remarriage in the United States. National Center for Health Statistics.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. Springer.

Brown, S. L. (2000). Union transitions among cohabitors: The significance of relationship assessment and expectations. Journal of Marriage and Family, 62(3), 833-846.

Bumpass, L. L., & Lu, H. H. (2000). Trends in cohabitation and implications for children's family contexts in the United States. Population Studies, 54(1), 29-41.

Cairo, A. (2012). The functional art: An introduction to information graphics and visualization. New Riders.

Carr, W., & Hudson, J. (2017). Open data for policy analysis: A practical guide. Bristol University Press.

Casper, L. M., & Bianchi, S. M. (2002). Continuity and change in the American family. Thousand Oaks, CA: Sage Publications.

Chang, K. (2007). R Graphics Cookbook. O'Reilly Media, Inc.

Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A taxonomy of open data portals. In Proceedings of the 17th Annual International Conference on Digital Government Research (pp. 328-333). Association for Computing Machinery.

Cherlin, A. J. (2004). The deinstitutionalization of American marriage. Journal of Marriage and Family, 66(4), 848-861.

Cherlin, A. J. (2009). The marriage-go-round: The state of marriage and the family in America today. Knopf.

Cleveland, W. S. (1993). Visualizing data. Hobart Press.

Dimmery, D. (2019). Machine learning for social scientists. Polity Press.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

Easterlin, R. A. (1980). Birth and fortune: The impact of numbers on personal welfare. Journal of Economic Literature, 18(4), 1603-1608.

Facebook. (2017). Prophet: Automatic Forecasting Procedure. https://facebook.github.io/prophet/

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research, 15(1), 3133-3181.

Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press.

Flach, P. A. (2012). Machine learning: The art and science of algorithms that make sense of data. Cambridge University Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York.

Goldscheider, F. K., & Waite, L. J. (1991). New families, no families? The transformation of the American home. Population and Development Review, 17(2), 327-333.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques. Morgan Kaufmann.

Harding, S. (1987). Feminism and methodology: Social science issues. Indiana University Press.

Haykin, S. S. (2009). Neural networks and learning machines. Pearson Education.

Healy, K. (2018). Data visualization: A practical introduction. Princeton University Press.

Heaton, T. B., & Blake, S. M. (1999). Gender differences in determinants of marital disruption. Journal of Family Issues, 20(6), 796-821.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer New York.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. Annual Review of Sociology, 24(1), 395-421.

Kane, E. W. (2022). The sociology of the family (2nd ed.). SAGE Publications, Inc.

Klein, D. M., & White, J. M. (2018). Theories of marriage and family (2nd ed.). SAGE Publications, Inc.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. MIT press.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer New York.

Lee, J. (2020). Unicef Malaysia: Child marriage likely rose during Covid-19 pandemic as schools closed, economy worsened. Malay Mail. Retrieved from https://www.malaymail.com/news/malaysia/2021/03/10/unicef-malaysia-child-marriage-likely-rose-during-covid-19-pandemic-as-scho/1956392

Lee, R. D. (1971). Population dynamics of humans and other animals. Demography, 8(3), 441-449.

Lesthaeghe, R., & van de Kaa, D. J. (1986). The second demographic transition in Western Europe: An interpretation. Population and Development Review, 12(3), 411–449.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lichter, D. T., & Qian, Z. (2008). Serial cohabitation and the marital life course. Journal of Marriage and Family, 70(4), 981-994.

MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

Manning, W. D., & Smock, P. J. (2005). Measuring and modeling cohabitation: New perspectives. Journal of Marriage and Family, 67(4), 989-1002.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill, Inc.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.

Morrow, R. A. (2013). Critical theory and methodology. Sage.

Murrell, P. (2011). R Graphics. Chapman and Hall/CRC.

Nielsen, M. A. (2015). Neural networks and deep learning. Determination press.

Oppenheimer, V. K. (1997). Women's employment and the gain to marriage: The specialization and trading model. Annual Review of Sociology, 23(1), 1 431-453.

Oppenheimer, V. K. (2003). A theory of marriage timing. American Journal of Sociology, 108(5), 1153-1193.

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, 2 Inc.

Qian, Z., & Lichter, D. T. (2007). Social boundaries and marital assimilation: Interpreting trends in racial and ethnic intermarriage. American Sociological Review, 3 72(1), 68-94.

Raley, R. K., & Bumpass, L. L. (2003). The topography of the divorce plateau: Levels and trends in union stability in the United States after 1980. Demographic Research, 4 8(7), 245-260.

Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer-Verlag New York.

Schoen, R., & Standish, N. J. (2001). Partner choice in marriages and cohabitations. Journal of Marriage and Family, 63(1), 51-63.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press. 5

Silver, N. (2012). The signal and the noise: Why so many predictions fail—but some don't. Penguin Press HC, The.

Smock, P. J. (2000). Cohabitation in the United States: An appraisal of research themes, findings, and implications. Annual Review of Sociology, 26(1), 6 1-20.

South, S. J. (1995). Racial and ethnic differences in the desire to marry. Journal of Marriage and Family, 57(1), 259-270.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Tan, L. (2019). Social impact of the COVID-19 pandemic in Malaysia. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Social_impact_of_the_COVID-19_pandemic_in_Malaysia

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Addison-Wesley Longman Publishing Co., Inc.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72.

Teachman, J. D. (2002). Stability across partnerships. Journal of Marriage and Family, 64(2), 280-291.

Thornton, A., & Young-DeMarco, L. (2001). Four decades of trends in attitudes toward family issues in the United States: The 1960s through the 1990s. Journal of Marriage and Family, 63(4), 1009-1037. 7

Tufte, E. R. (2001). The visual display of quantitative information. Graphics Press.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

Vapnik, V. N. (2013). The nature of statistical learning theory. Springer science & business media.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 8 3-28.

Waite, L. J., & Gallagher, M. (2000). The case for marriage: Why married people are happier, healthier, and better off financially. New York: Doubleday. 9

Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3-28.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wilke, C. O. (2019). Fundamentals of data visualization: A primer on making informative and compelling figures. O'Reilly Media, Inc. 10

Wilkinson, L. (2005). The grammar of graphics. Springer Science & Business Media.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 11

Wu, Z., & Pollard, M. S. (2000). Economic circumstances and the stability of nonmarital cohabitation. Journal of Marriage and Family, 62(3), 623-636.

Yau, N. (2011). Visualize this: The flowing data guide to design, visualization, and statistics. John Wiley & Sons.

Zuiderwijk, A., Janssen, M., & Choenni, S. (2012). The value of open data. Journal of Theoretical and Applied Electronic Commerce Research, 7(1), 1-17.