SC1015

Mini Project

ZOE TAN XUAN YO
MAI XIWEN
LUX PANG JIANWEN

Content

- Introduction
- Problem Statement
- Data Preparation
- Basic Visualization
- Exploratory Data Analysis
- Model Building & Prediction
- Conclusion and Insights

Introduction

- Low fertility rates can lead to negative consequences such as aging population, economic decline, and social challenges
- A shrinking population can affect political and social stability, and change social norms and cultural values
- Governments should address this issue by implementing policies and programs that support families and child-rearing

Problem Statement

We aim to investigate the factors that contribute to declining fertility rates in Singapore and predict the fertility rate for the next decade.

We will explore the relationships between education level, cost of living, rate of marriage, and fertility rates to help address the issue of an aging population and economic challenges.

Part 1:
Data Preparation

Steps in Data Preparation

Step 1:
Renaming columns

Step 2:
Set Index to Years

Step 3:
Preliminary Feature Selection

Step 4:
Dealing with Null Values

Step 5:
Converting DataFrames to Pickle Files

```
easier reading
['Data Series': 'Years'})

['TotalFertilityRate(PerFemale)': 'TotalFertilityRate'})
['15to19Years(PerThousandFemales)': 'F.15to19Years'})
['20to24Years(PerThousandFemales)': 'F.20to24Years'})
['25to29Years(PerThousandFemales)': 'F.25to29Years'})
['30to34Years(PerThousandFemales)': 'F.30to34Years'})
['35to39Years(PerThousandFemales)': 'F.35to39Years'})
['40to44Years(PerThousandFemales)': 'F.40to44Years'})
['45to49Years(PerThousandFemales)': 'F.45to49Years'})


['TotalFemalestoBelowSecondary': 'BelowSecondary'})
['TotalFemalestoSecondary': 'Secondary'})
['TotalFemalestoPostSecondary(NontoTertiary)': 'PostSecondary(NontoTert
['TotalFemalestoDiploma&ProfessionalQualification': 'Diploma&Professiona
['TotalFemalestoUniversity': 'University'})


['FemaleGeneralMarriageRate(Per1000UnmarriedResidentMalesOrFemalesAged15
['15to19Years(Per1000UnmarriedResidentFemales)': 'M.15to19Years'})
['20to24Years(Per1000UnmarriedResidentFemales)': 'M.20to24Years'})
['25to29Years(Per1000UnmarriedResidentFemales)': 'M.25to29Years'})
['30to34Years(Per1000UnmarriedResidentFemales)': 'M.30to34Years'})
['35to39Years(Per1000UnmarriedResidentFemales)': 'M.35to39Years'})
['40to44Years(Per1000UnmarriedResidentFemales)': 'M.40to44Years'})
['45to49Years(Per1000UnmarriedResidentFemales)': 'M.45to49Years'})
['50to54Years(Per1000UnmarriedResidentFemales)': 'M.50to54Years'})
['55to59Years(Per1000UnmarriedResidentFemales)': 'M.55to59Years'})
['60to64Years(Per1000UnmarriedResidentFemales)': 'M.60to64Years'})
['65Years&Over(Per1000UnmarriedResidentFemales)': 'M.65Years&Over'})
```

Step 1

Renaming Columns

Key features

- We renamed 'Data Series' to 'Years' for clarity
- Removed '(Per Thousand Females)' in the columns names and state it once in our dataset.
- Removed 'Total Females' in the columns names and state it once in our dataset.

Years	TotalFertilityRate	F.15to19Years	F.20to24Years
2022	1.05	2.1	11.3
2021	1.12	2.2	11.7
2020	1.10	2.3	12.7
2019	1.14	2.5	12.7
2018	1.14	2.5	14.4
...	...	...	...
1965	4.66	35.9	227.1
1964	4.97	38.3	240.0
1963	5.16	45.7	249.0

Step 2

Set Index to Years

Key features

- Organizing fertility rate data by setting the index to years helps to identify trends and patterns over time.
- Using time-based functions and operations enables calculation of annual averages and comparisons with other variables.
- Time-based charts can provide a clear visualization of changes and patterns over time.
- This approach can aid in understanding how fertility rates have changed over time and identifying potential influencing factors.

Step 3

Preliminary Feature Selection

To manage our dataset's many variables, we opted to work with a carefully selected subset.

The choice of the variables has been made carefully. Since our question relates to the Singapore's declining fertilty rate and the factors that caused it, the variables we have chosen are:

Variables Relating to Fertility(Per 1000 Females):

- TotalFertilityRate(Per female)
- 15to19Years
- 20to24Years
- 25to29Years
- 30to34Years
- 35to39Years
- 40to44Years
- 45to49Years
- 50to54Years
- 55to59Years
- 60to64Years
- 65Years&Over

Variables Relating to Education:

- BelowSecondary
- Secondary
- PostSecondary(NontoTertiary)
- Diploma&ProfessionalQualification
- University

Variables Relating to Cost of Living:

- AllItems
- Food
- Clothing&Footwear
- Housing&Utilities
- HouseholdDurables&Services
- HouseholdServices&Supplies
- HealthCare
- Transport
- PrivateTransport
- PublicTransport
- OtherTransportServices
- Communication
- Recreation&Culture
- Education
- MiscellaneousGoods&Services
- PersonalCare
- AlcoholicDrinks&Tobacco
- PersonalEffects
- AllItemsLessImputedRentalsOnOwnertoOccupiedAccommodation
- AllItemsLessAccommodation

Variables Relating to Marriage(Per1000UnmarriedResidentFemales):

- FemaleGeneralMarriageRate
- 15to19Years
- 20to24Years
- 25to29Years
- 30to34Years
- 35to39Years
- 40to44Years
- 45to49Years
- 50to54Years
- 55to59Years
- 60to64Years
- 65Years&Over

Step 4

Dealing with Null Values

For more accurate and consistency analysis we have removed the null values by replacing them with the mean values of the column.

In [89]:

```
#Change all the 0s to mean value
EduFactor = EduFactor.replace(0, EduFactor.mean())
CoLFactor = CoLFactor.replace(0, CoLFactor.mean())
MarFactor = MarFactor.replace(0, CoLFactor.mean())
FerFactor = FerFactor.replace(0, CoLFactor.mean())
```

Step 5

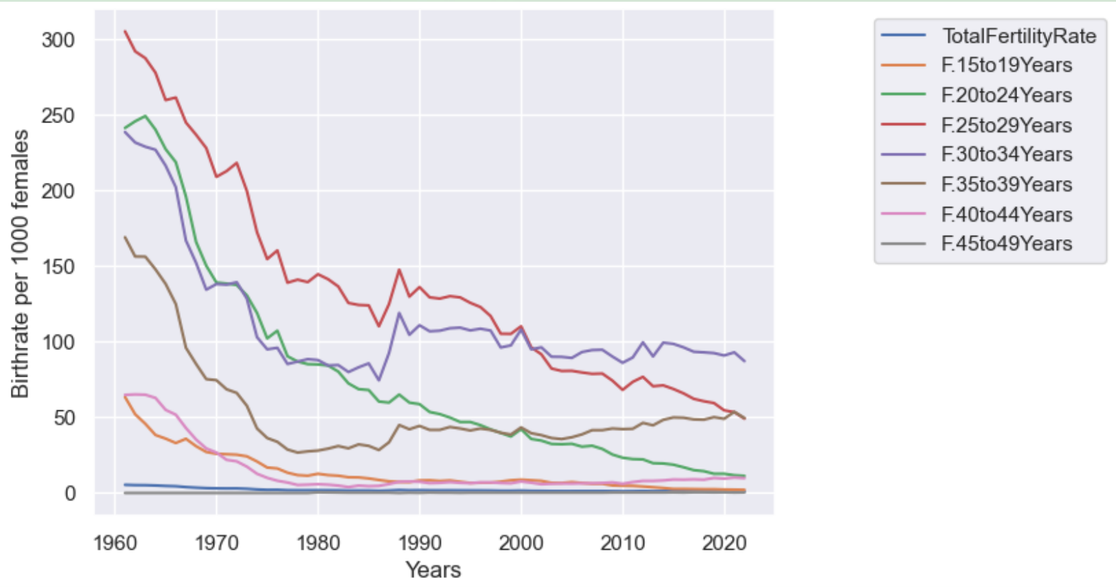
Converting DataFrames to Pickle Files

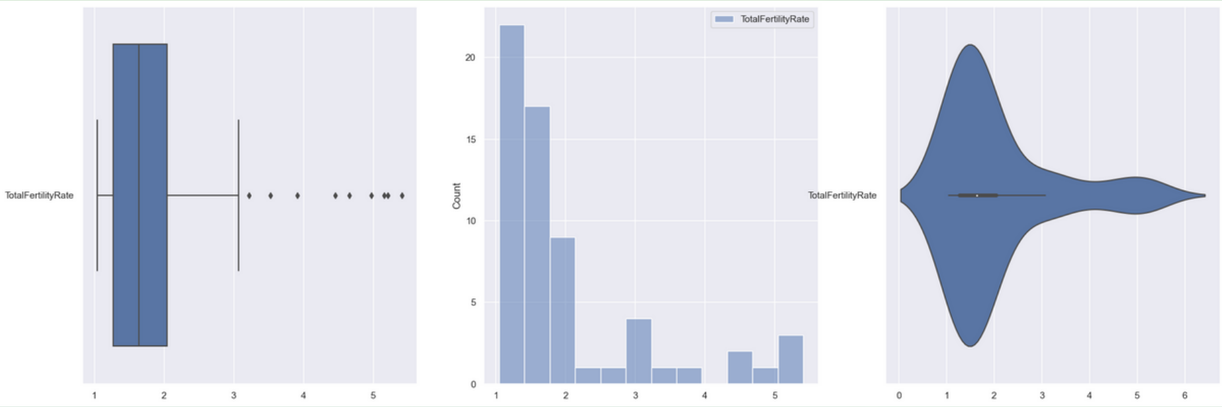
Finally, we convert our dataframes to Pickle files so that we are able to collaborate easily and do the rest of the project smoothly.

0]:

```
#Saving the data as picklefiles
EduFactor.to_pickle('EduFactor.pickle')
CoLFactor.to_pickle('CoLFactor.pickle')
MarFactor.to_pickle('MarFactor.pickle')
FerFactor.to_pickle('FerFactor.pickle')
```

Part 2:
Basic Visualisation



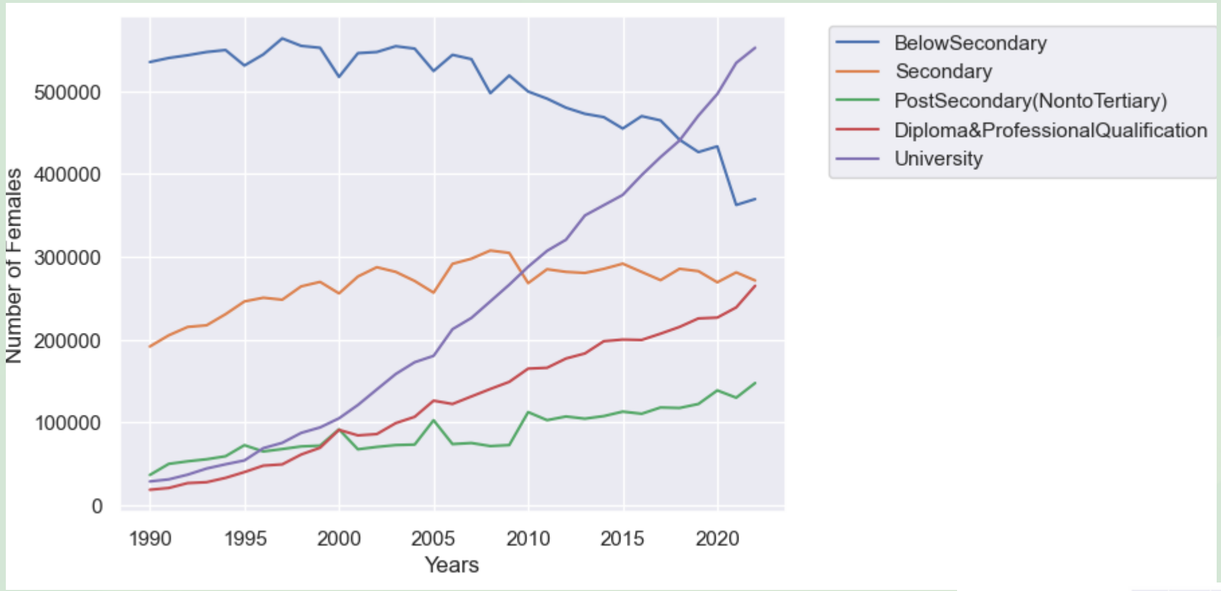


Fertility Rate Basic Visualisation

- We used Box plot, bar graph, violin plot and line graph to get some insights of the dataset
- However box plot and violin plot may not be useful for visualizing fertility rate data as it is typically normally distributed and reported at the country level.

Insight #1 from EDA and Basic Visualisation

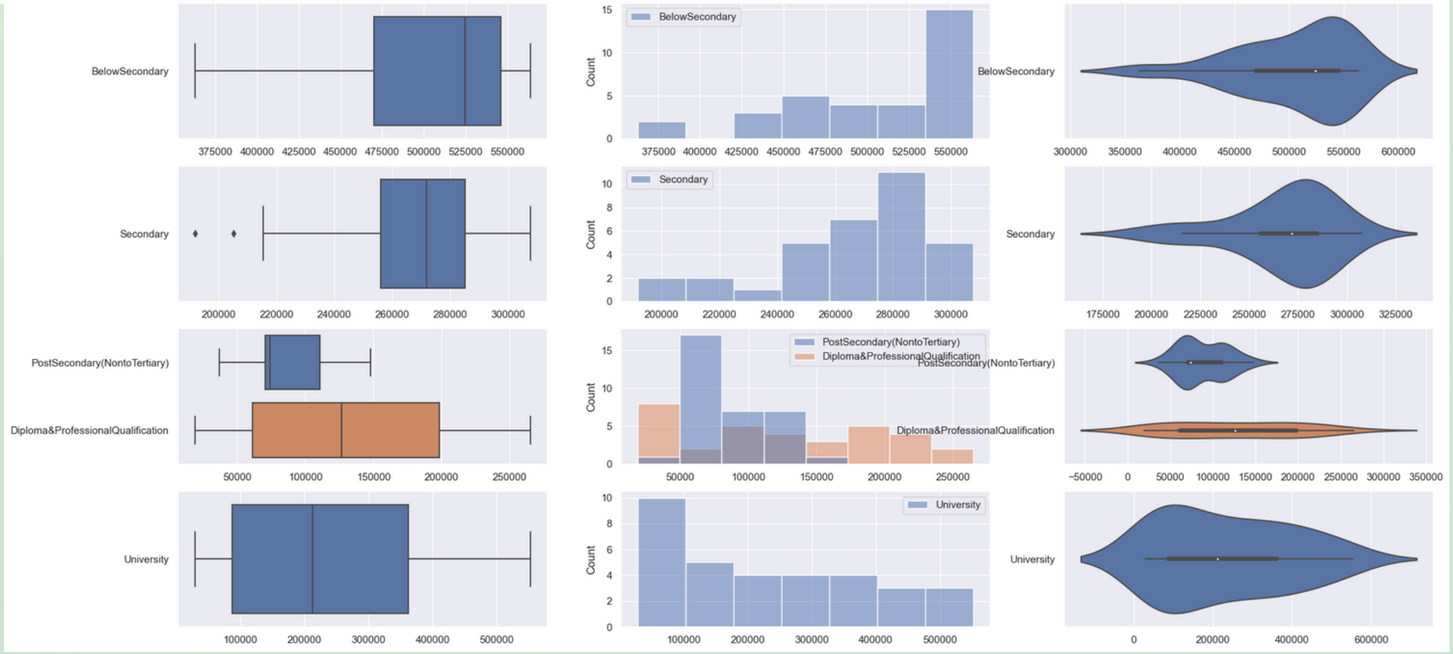
- Based on the data analysis, we can see that the overall trend in fertility rates has been declining over the years.
- However, there has been an interesting and somewhat unexpected increase in fertility rates among women aged above 30 since the 1990s.
- This increase in fertility rates could be due to various factors, such as the rise in education levels among women and a shift towards delaying childbirth until later in life.

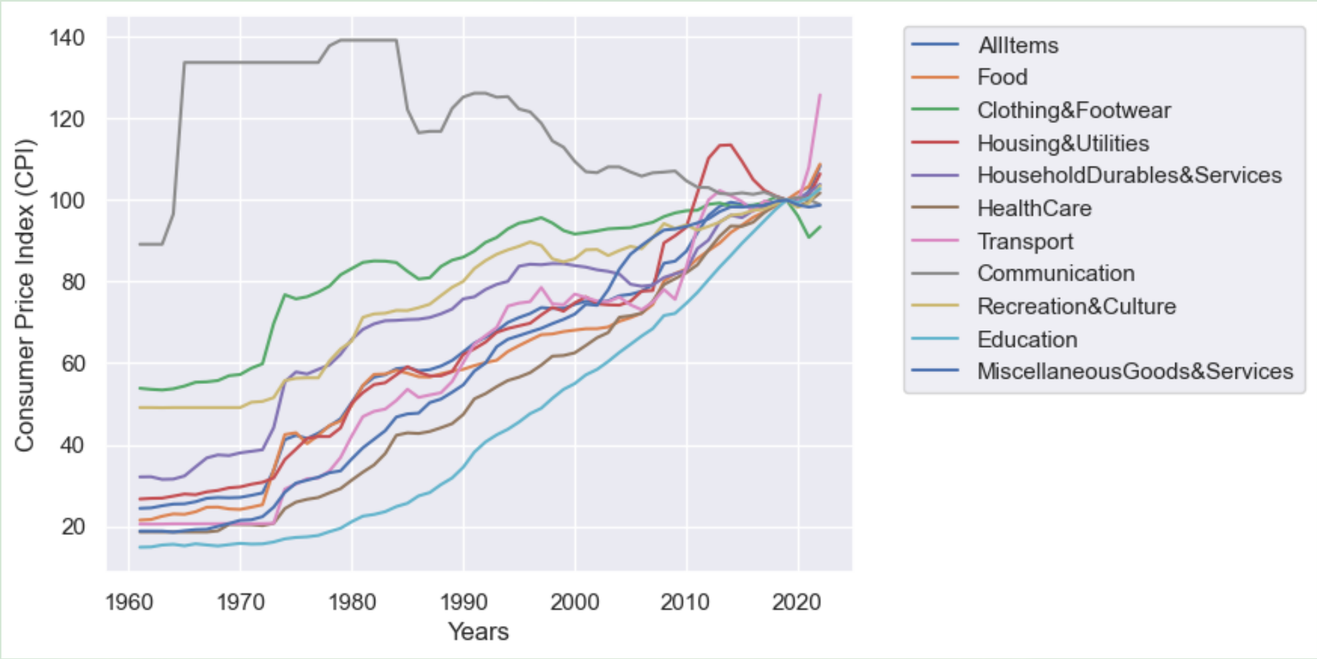


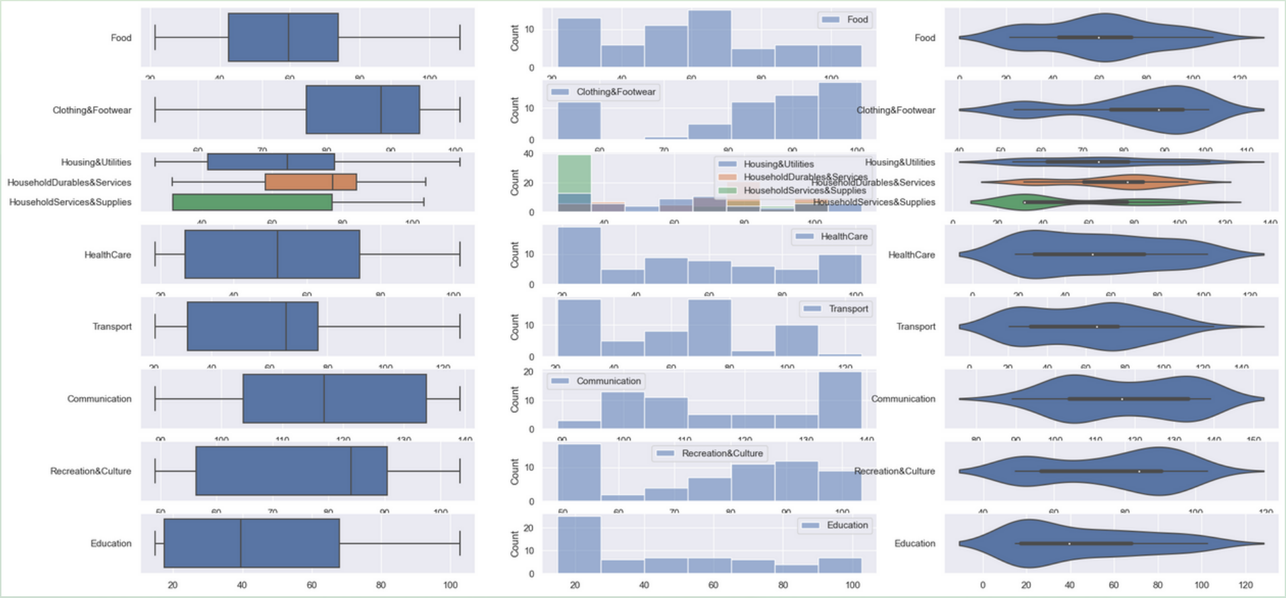
Education Basic Visualisation

From the line graph we can tell that there is a general trend of increasing in education for females in Singapore through out the years.

The graph shows that females who enrol and graduate from university increase significantly since 1990.





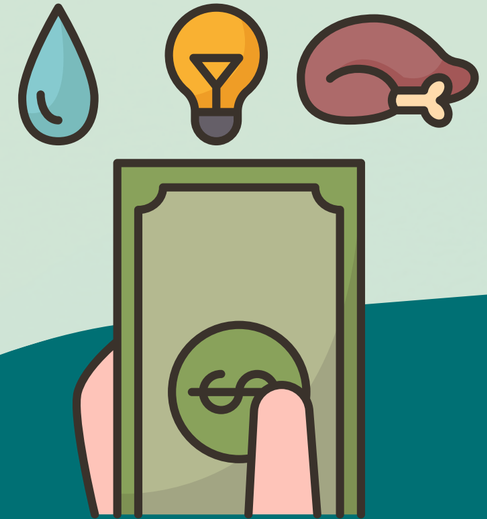


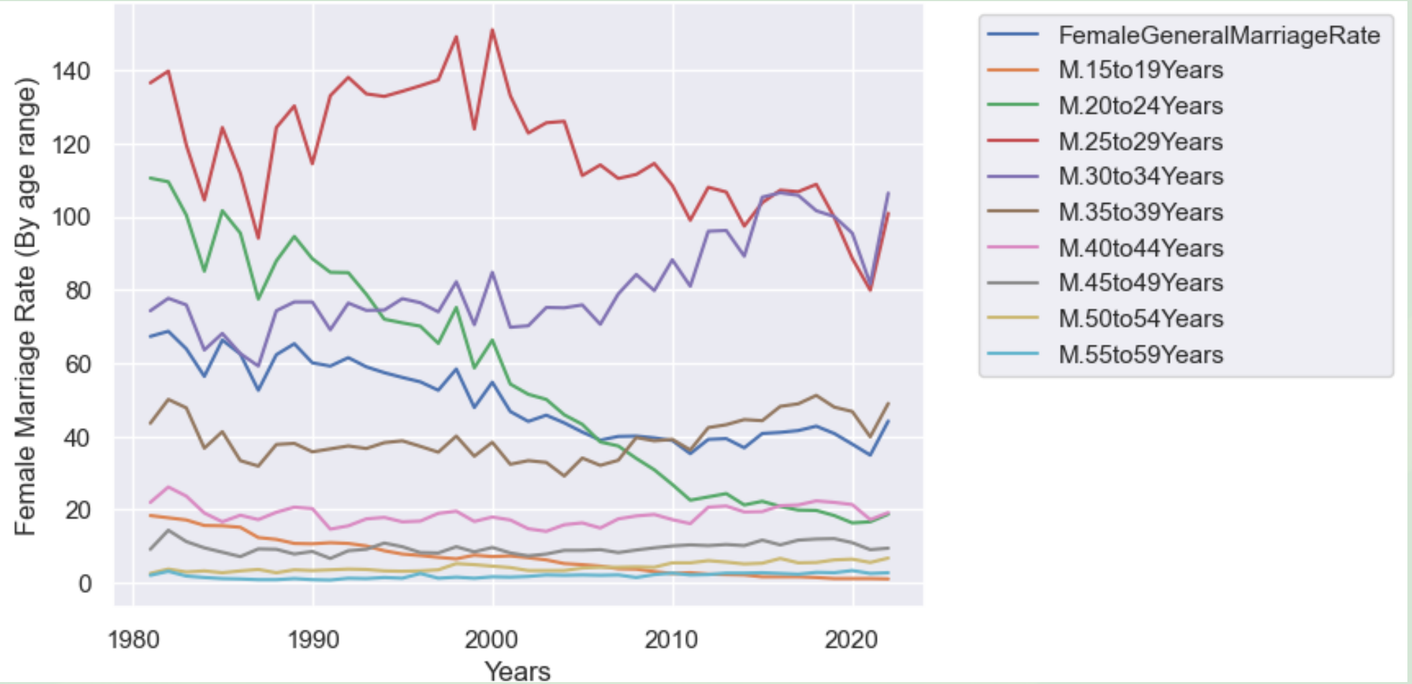
Cost of Living Basic Visualisation

The cost of living factor is measured by Consumer Price Index (CPI) and we can see the over the years all of the itmes in Singapore is increasing.

The CPI in Singapore has generally trended upwards over the years, reflecting the increase in prices of goods and services consumed by households in the country.

This reflects the rising costs of housing, healthcare, and education, among other factors.

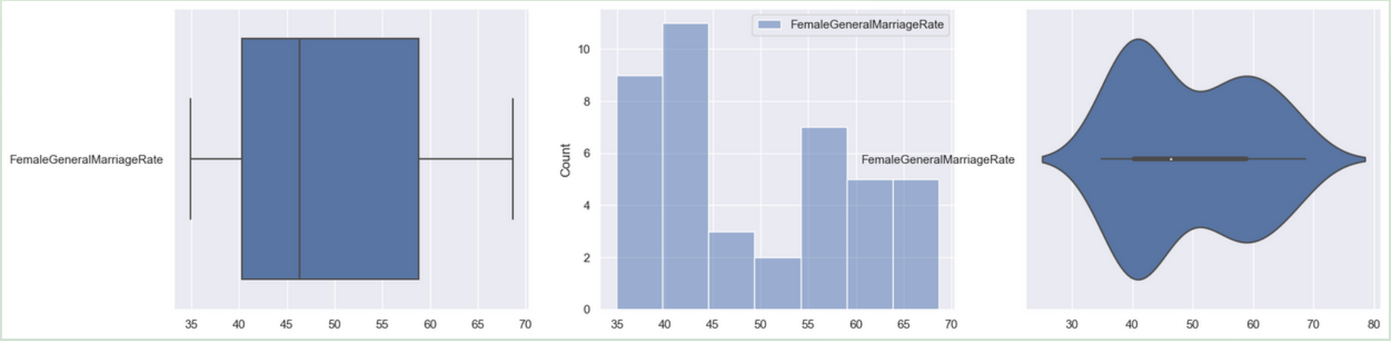




Marriage Rate Factor Basic Visualisation

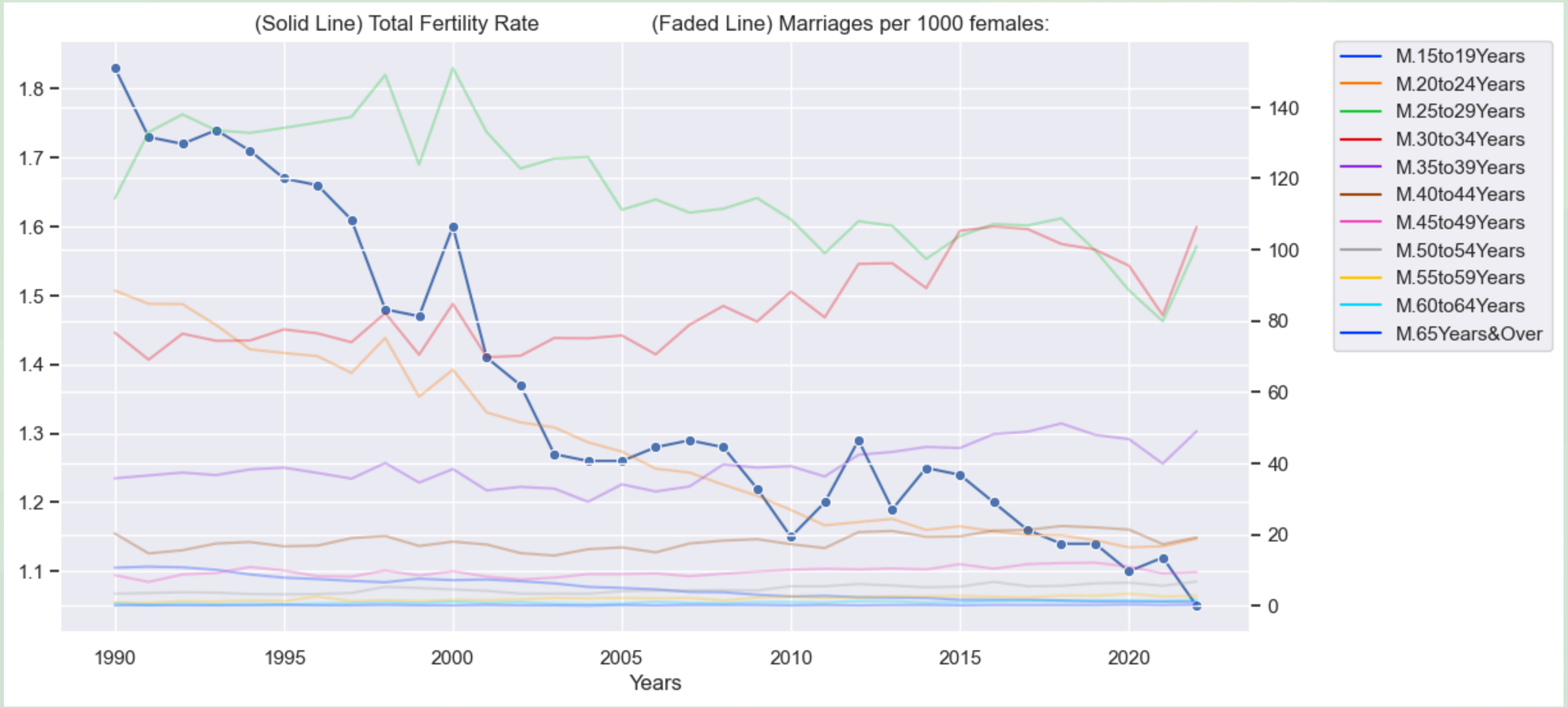
The marriage rate for females in Singapore has declined over the years, reflecting changes in societal attitudes towards marriage and family.

We can observed from the graph that especially Marriage at 20 to 24 years has been declining significantly over the years

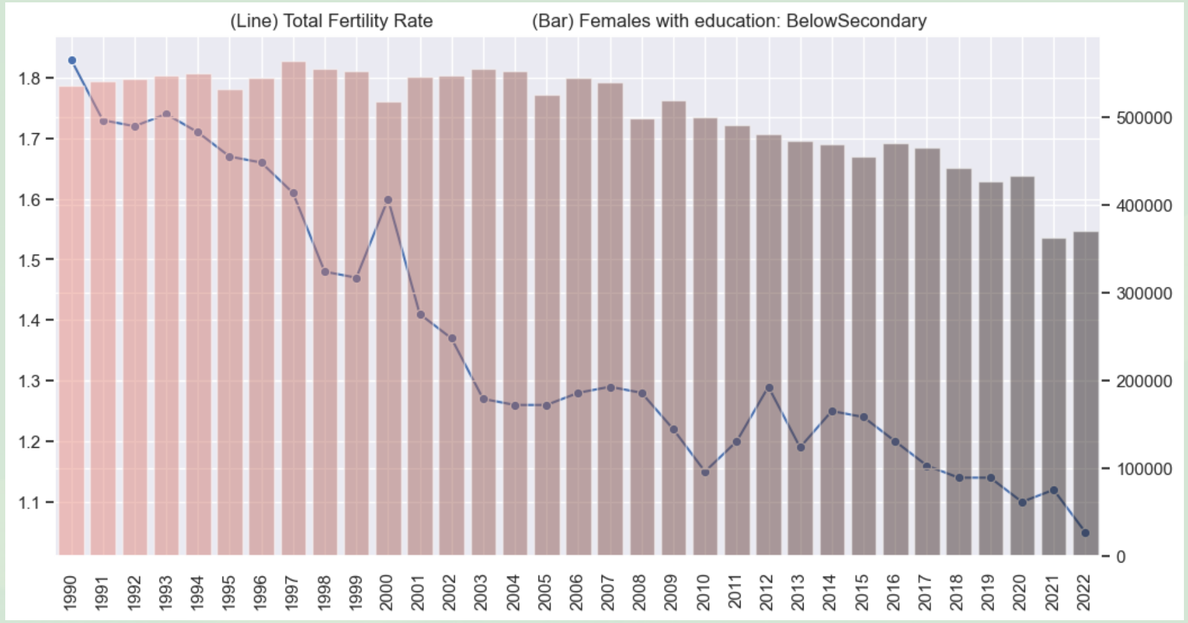


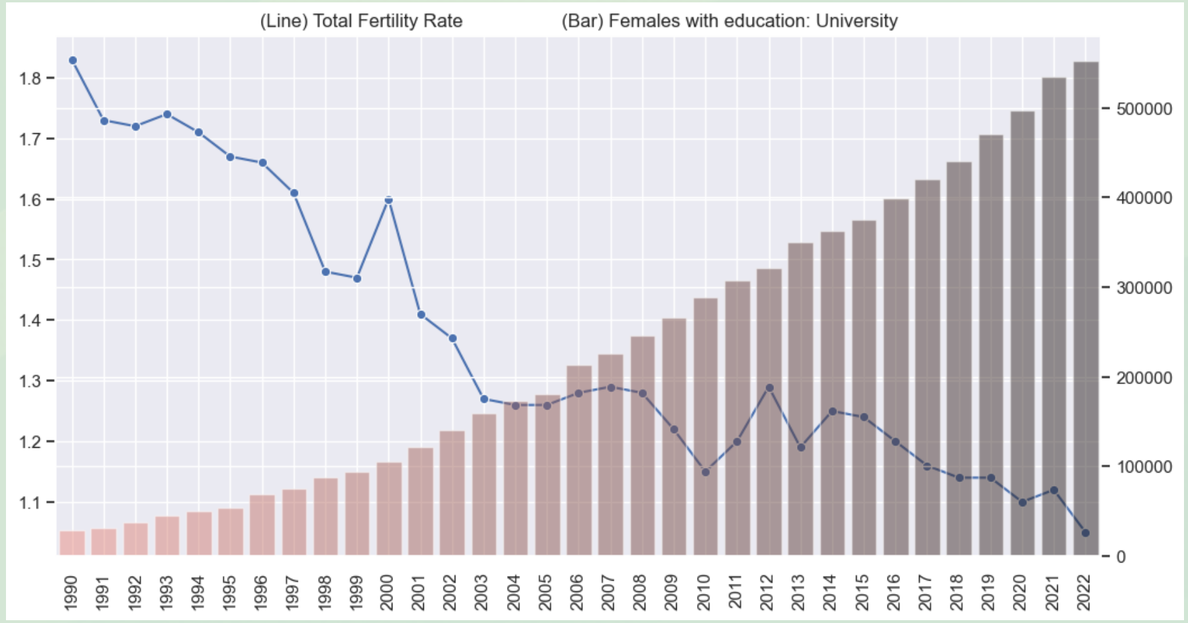
Part 3:
Exploratory Data Analysis

Marriage Rate in Females Factor Vs Fertility Rate



Education Factor Vs Fertility Rate

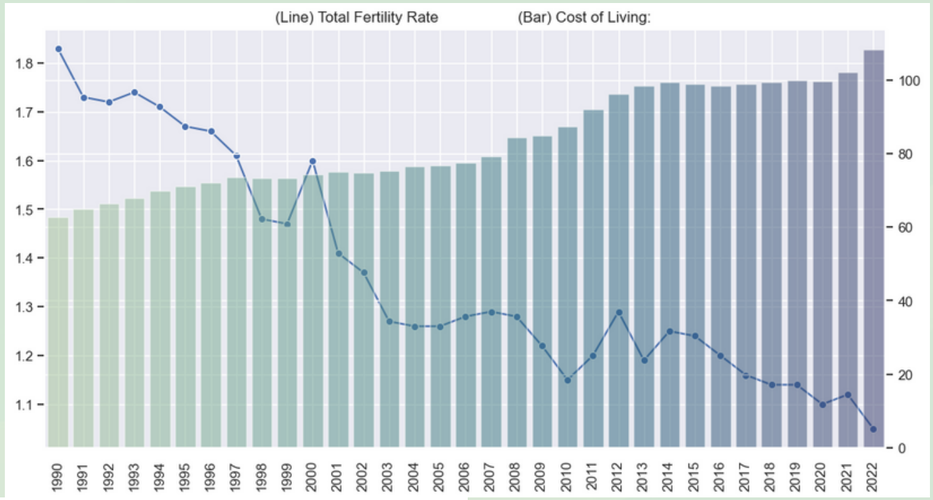


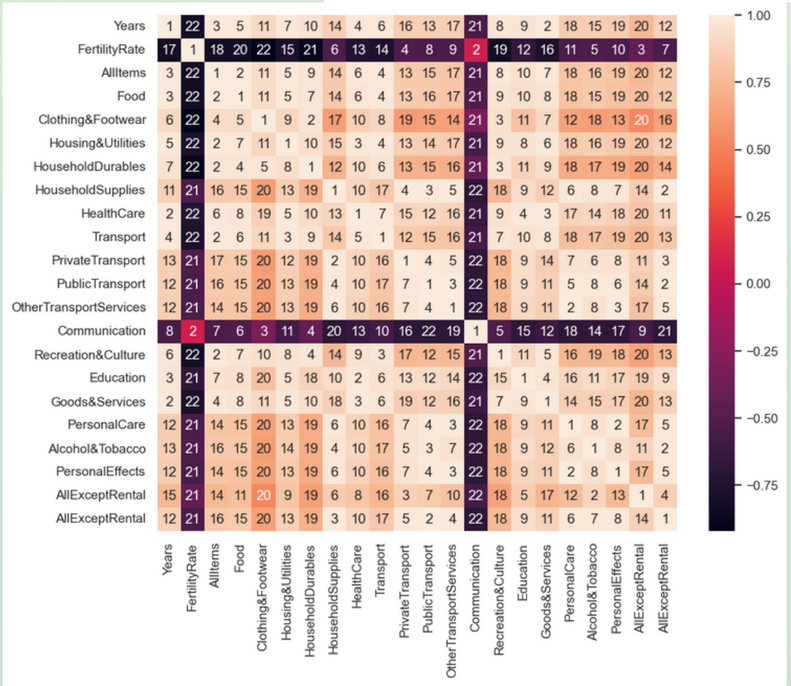


Insight #2 from EDA and Basic Visualisation

- We see a general upward trend in highest qualification attained by women especially in University. We can see that in 1990 there are about 2,500 females who graduated in University as compared to 2022 with more than 500,000 females who graduated from University.
- The increase in education levels among women could have had an impact on their decision to delay childbirth until after they have completed their education and established their careers.
- This trend could also be influenced by changes in societal norms and expectations surrounding marriage and motherhood.

Cost of Living Factor Vs Fertility Rate

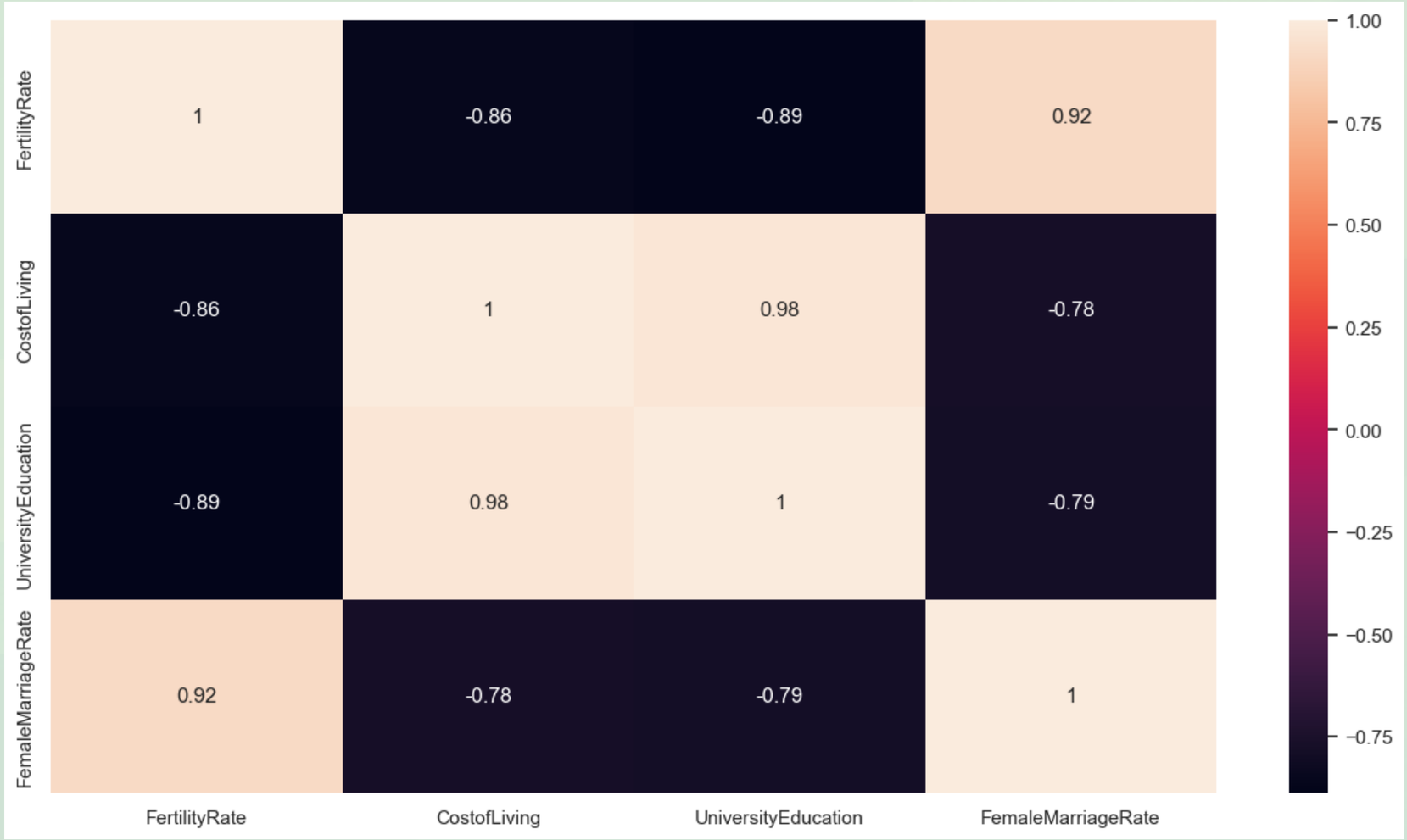




Insight #3 from EDA and Basic Visualisation

- The visualization suggests that the cost of living, with the exception of communication costs, has a negative correlation with fertility rates.
- Shows that while cost of living increases, fertility rates tend to decrease.
- This finding highlights the financial burden that comes with starting a family and the importance of affordable childcare, housing, and other necessary expenses for young families.

Correlation Matrix for all factors and the Fertility rate in Singapore



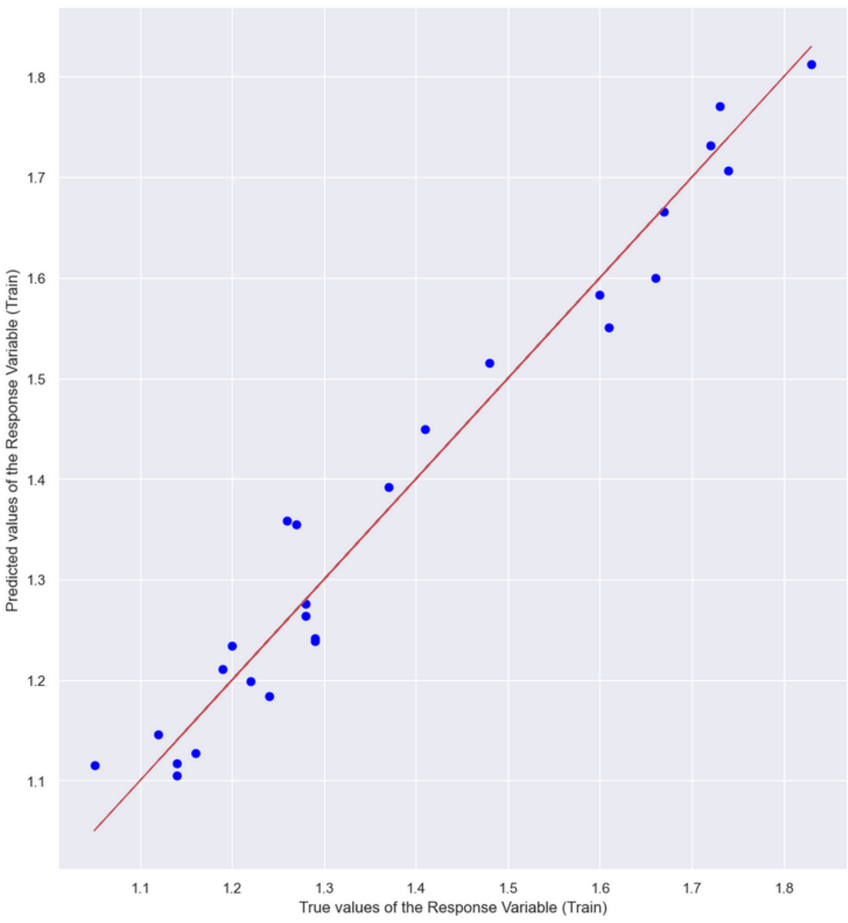
Part 4: Prediction & Modelling

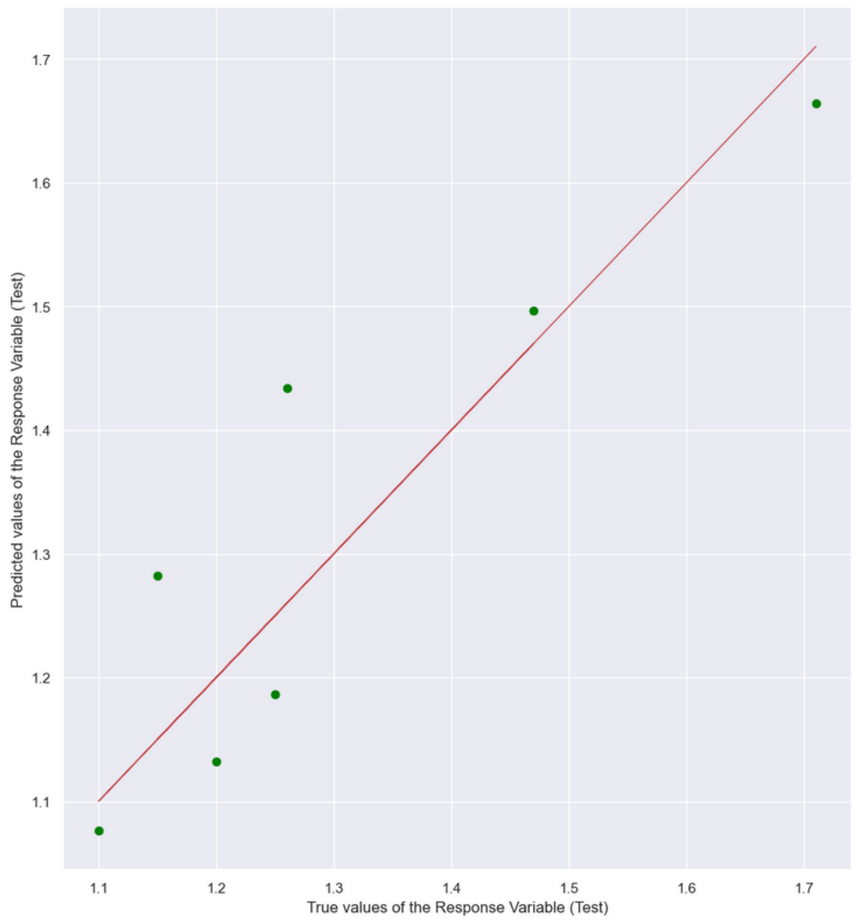
Multi-Variate Linear Regression

- We opted for Multi-Variate Linear Regression over Uni-Variate Linear Regression.
- While testing the Education dataset, we discovered that 'University' level of education was the strongest predictor among all variables.
- However, this information was too specific and lacked value as we aimed to predict the general fertility rate.
- Thus, a Multi-Variate Linear Regression model will help determine if 'Education', 'Cost of Living', or 'Marriage Rates' is a good predictor for 'Total Fertility Rate' in general.

Using Education to predict Total Fertility Rate

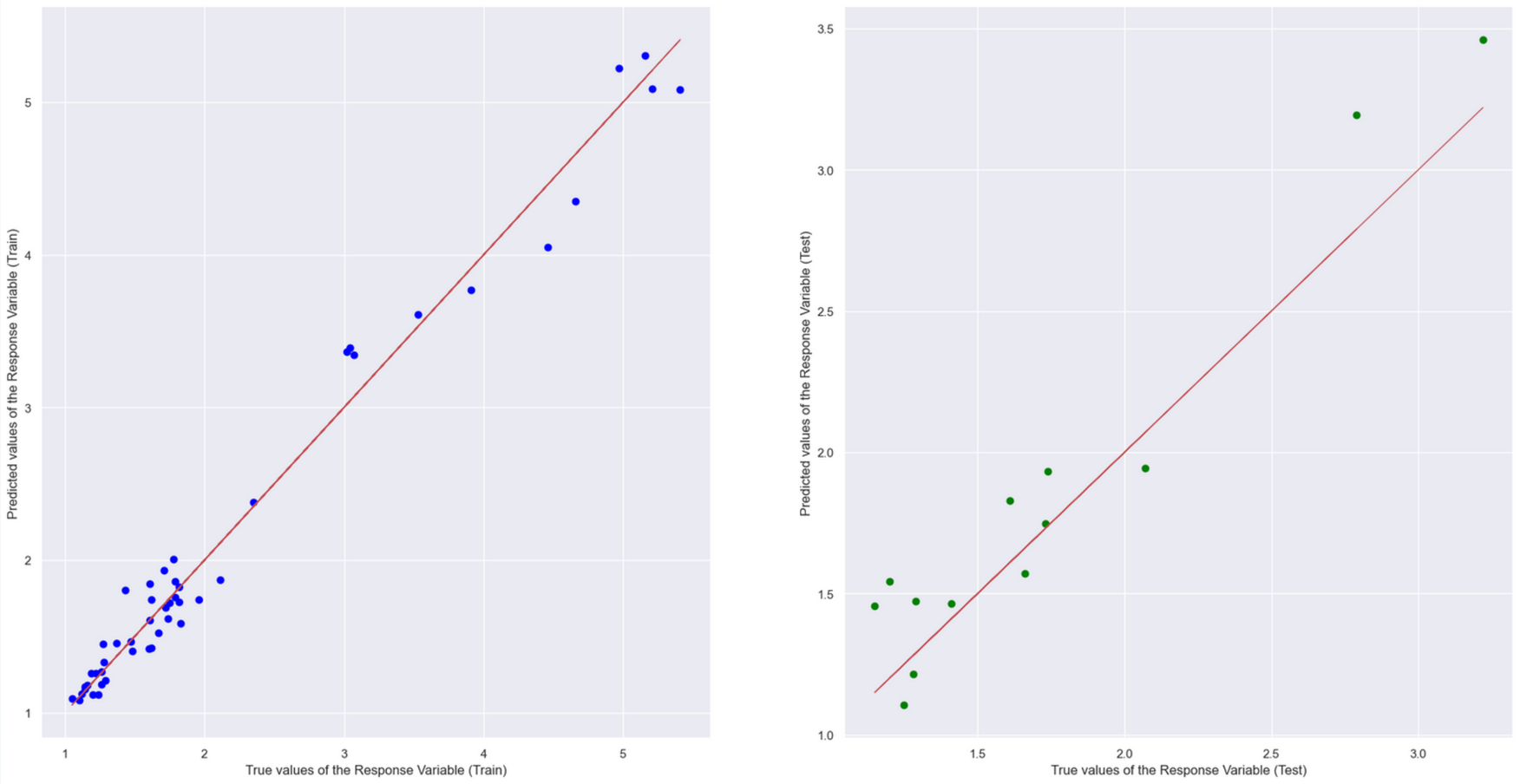
- Scatter plot for TotalFertilityRate in education dataset shows positive linear relationship with education levels
- Higher education levels offer better access to family planning, influencing demographic patterns
- Scatter plot provides valuable starting point for further analysis despite outliers and variability in test set.





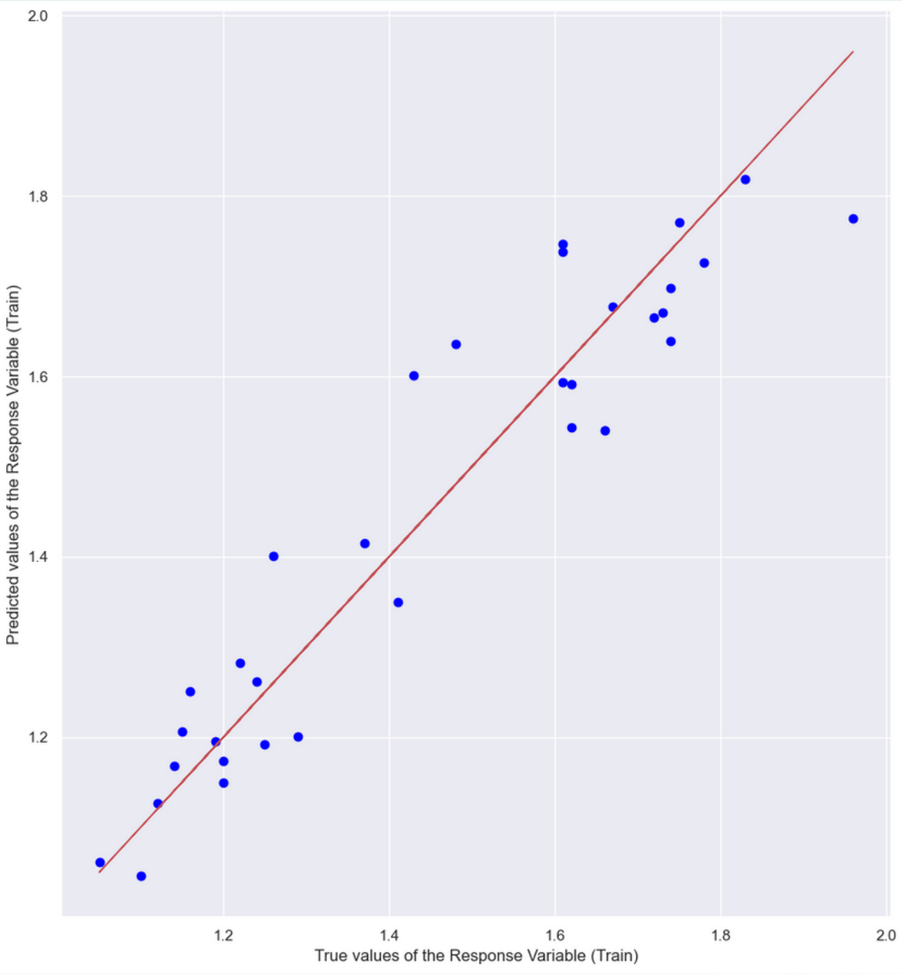
Using Cost of Living to predict Total Fertility Rate

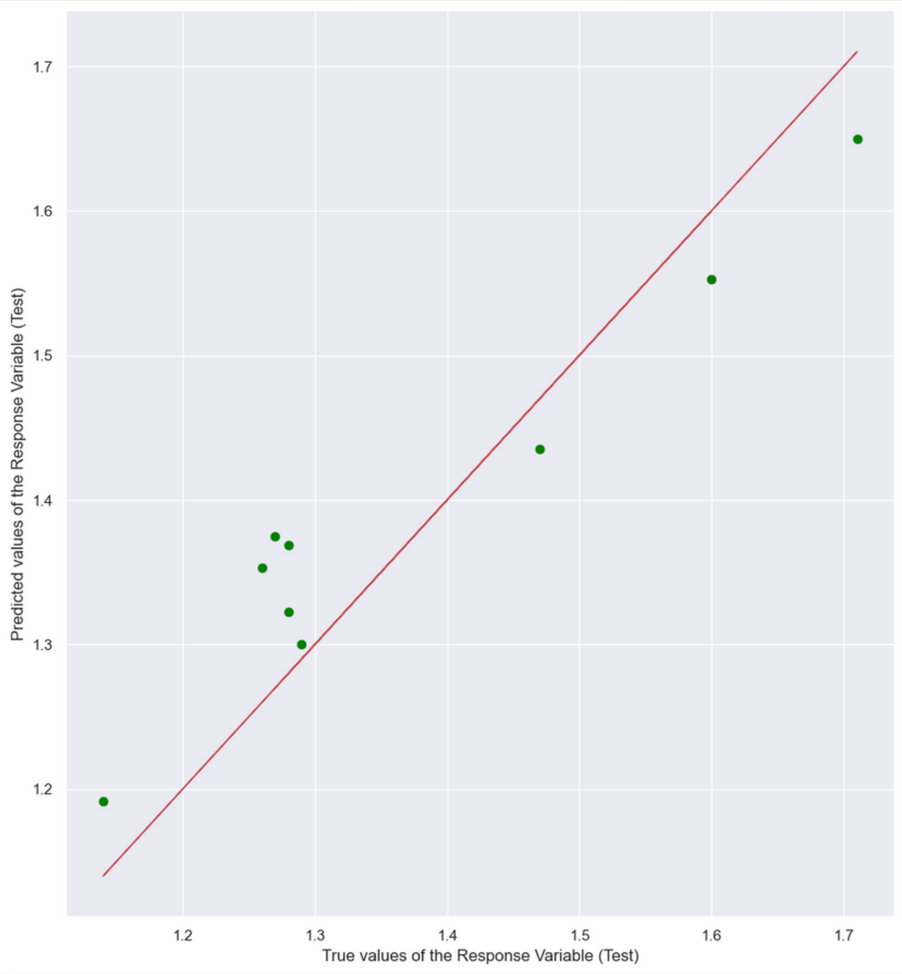
- Scatter plots in marriage rates dataset suggest weak correlation between true and predicted TotalFertilityRate values for both training and testing sets.
- Large deviation from ideal scenario, indicating poor accuracy in model predictions.
- Model may require further refinement to improve its predictive power.
- Weak correlation suggests that other variables may be more important predictors of TotalFertilityRate.



Using Marriage Rate to predict Total Fertility Rate

- Linear regression model provides a good fit for TotalFertilityRate in marriage dataset for both training and test sets.
- Selected predictor variables (FemaleGeneralMarriageRate and age-specific marriage rates) accurately predict TotalFertilityRate.
- Potential for underfitting as the model assumes a linear relationship between predictors and response variable.
- Linear regression may miss some variations in the data.





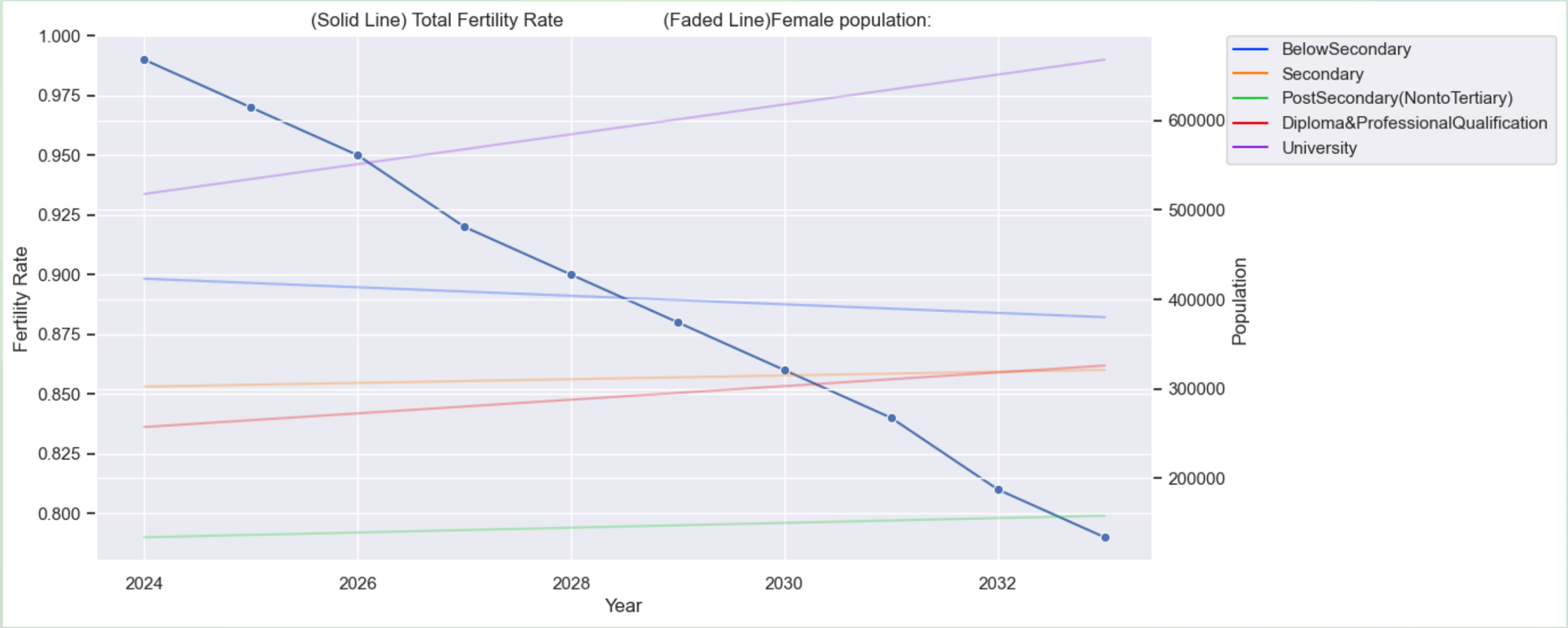
Conclusion

- Education has high R^2 and low MSE on both Train and Test sets, indicating a well-fitted model with accurate predictions for new data.
- Cost of Living has high R^2 on Train set, but high MSE on Test set, suggesting overfitting on Train set and poor generalization.
- Marriage Rates has low R^2 on both Train and Test sets, but low MSE on both sets, indicating underfitting and incomplete data capture.
- Education is the most accurate predictor with high R^2 and low MSE on both Train and Test sets.

Prediction of Fertility Rate for the next Decade

	TotalFertilityRate	BelowSecondary	Secondary	PostSecondary(NontoTertiary)	Diploma&ProfessionalQualification	University
2024	0.99	422884.10	302235.40	133821.20	257090.20	517528.60
2025	0.97	418103.50	304312.20	136504.60	264719.70	534213.00
2026	0.95	413322.90	306389.00	139188.10	272349.20	550897.40
2027	0.92	408542.20	308465.70	141871.50	279978.80	567581.80
2028	0.90	403761.60	310542.50	144554.90	287608.30	584266.20
2029	0.88	398981.00	312619.30	147238.30	295237.90	600950.60
2030	0.86	394200.40	314696.00	149921.70	302867.40	617634.90
2031	0.84	389419.80	316772.80	152605.20	310497.00	634319.30
2032	0.81	384639.20	318849.60	155288.60	318126.50	651003.70
2033	0.79	379858.60	320926.40	157972.00	325756.10	667688.10

Prediction of Fertility Rate for the next Decade



Prediction of Fertility Rate for the next Decade

- Based on the above findings, the predicted Total Fertility Rate for the next 10 years is on a straight decline, with the Correlation Coefficient being at -0.999, it is almost a Perfect Negative correlation.

Thank you!