Scot Forge Data Analysis Project

Submitted by: Group 10

Chandan Rudrappa Mukesh Ratakonda Nandani Rabra

Table of Contents

Exploratory Data Analysis (EDA)

1. Overview of the Dataset

The dataset has 1,243 entries and 9 columns. Here is what each column represents:

- **children**: Number of children each woman has.
- **german**: Whether the woman identifies as German.
- **years_school**: The number of years spent in school.
- **voc_train**: If the woman attended vocational training.
- **university**: If the woman attended university.
- **religion**: The woman's religion.
- **year_birth**: The year the woman was born.
- **rural**: Whether the woman lives in a rural area.
- **age_marriage**: The age when the woman got married.

Data Quality

- There are no missing values in the dataset.
- The data types are correct (e.g., numbers for numerical columns, text for categories).

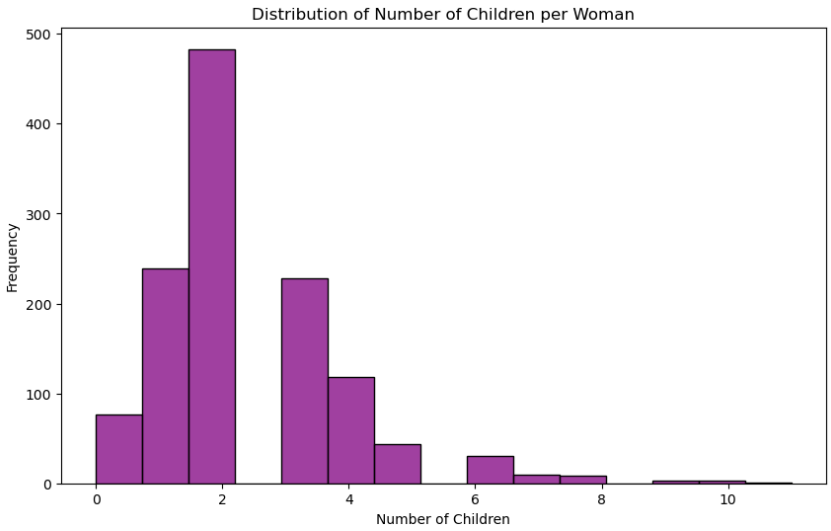
Summary of the Data

- **Children**: On average, women have 2.38 children. The variance is 2.33.
- **Years in School**: The average is 9.1 years, with a standard deviation of 0.95.
- **Year of Birth**: Women were born between 1940 and 1983.
- **Age at Marriage**: The average age is 23, ranging from 17 to 30

2. Key Insights from Graphs

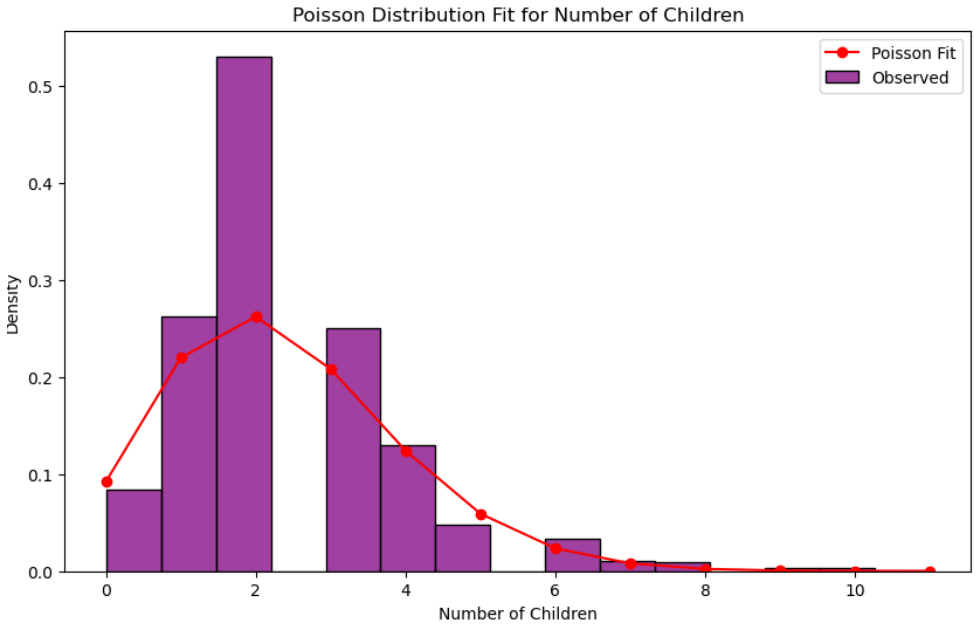
a) Distribution of Number of Children

- **What the Graph Shows**: A histogram of how many children each woman has.
- **Key Point**: Most women have 2-3 children. The data has a slight skew to the right.



b) Poisson Distribution Fit

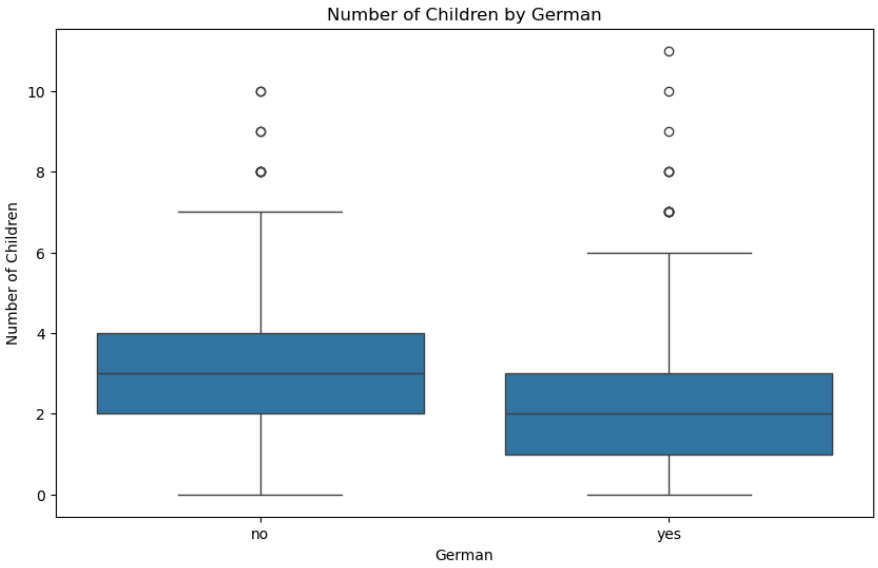
- **What the Graph Shows**: The histogram is compared to a Poisson distribution curve.
- **Key Point**: The mean (2.38) and variance (2.33) are close, which matches a Poisson distribution.



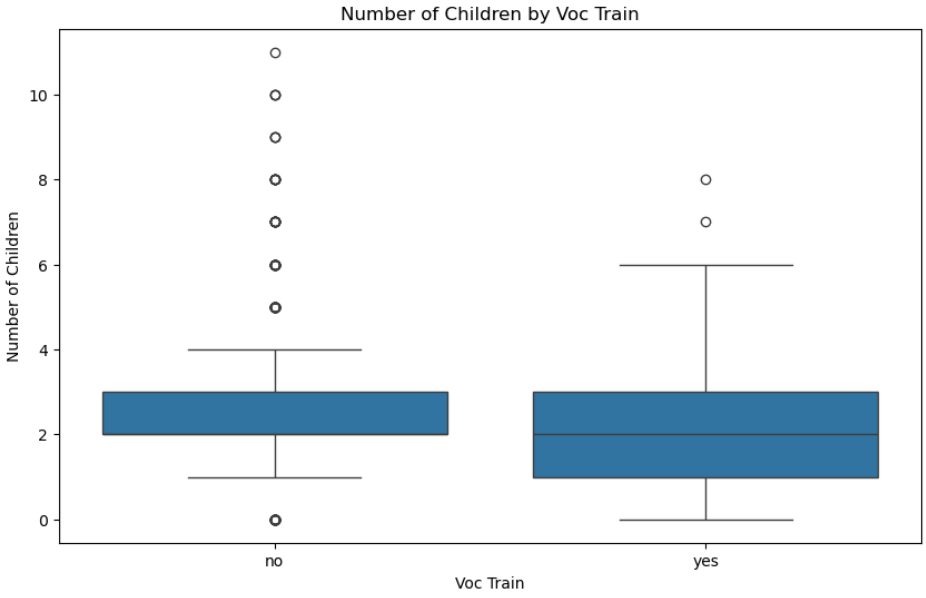
c) Number of Children by Categories

Boxplots show how "children" varies for different groups:

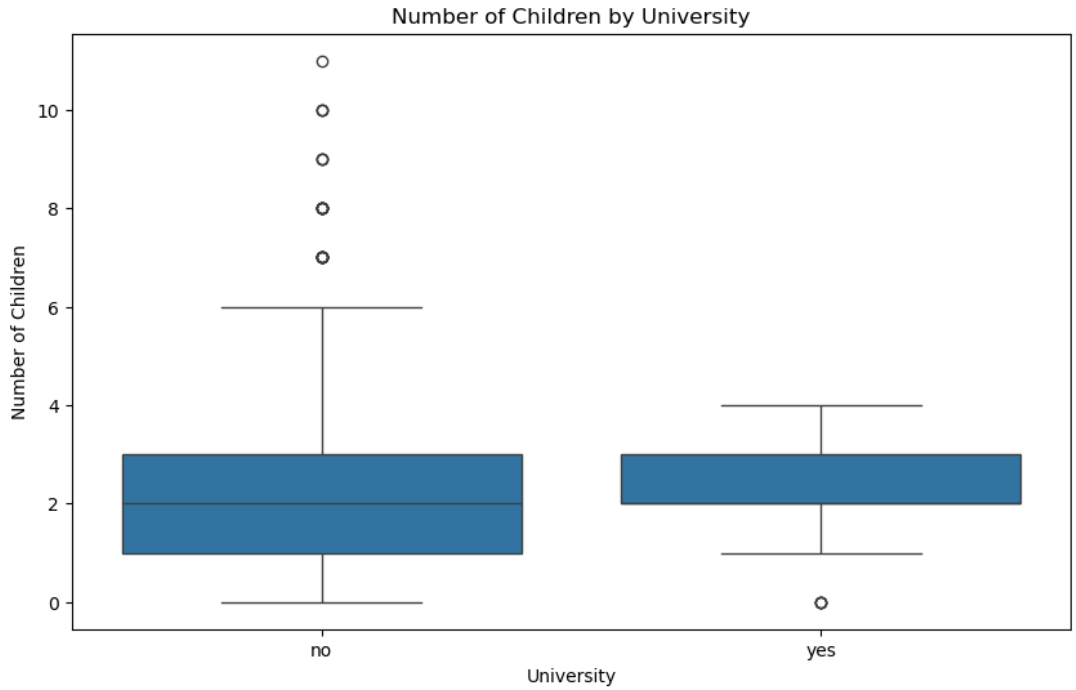
- **German**: No major differences.
- **Vocational Training**: Women with training have slightly fewer children.
- **University**: Women who went to university have fewer children.
- **Religion**: Small variations among religious groups.
- **Rural vs Urban**: Women in rural areas have more children.



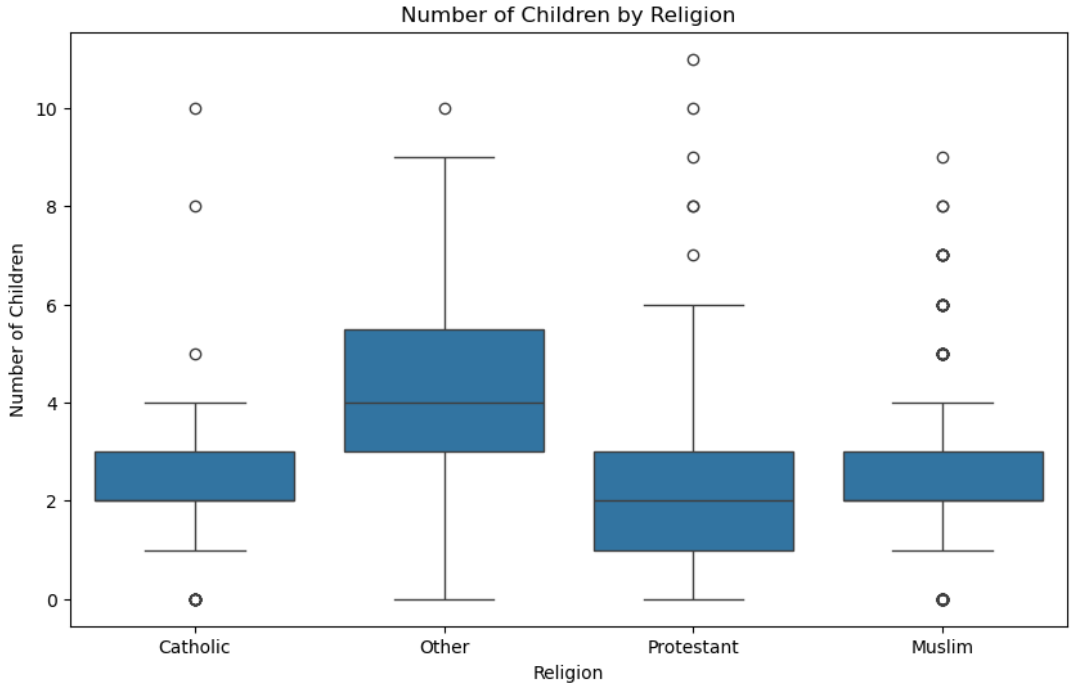

```
Chi-square test for german:
Chi-square Statistic: 0.1795018706475723, p-value: 0.6718016749249864
```



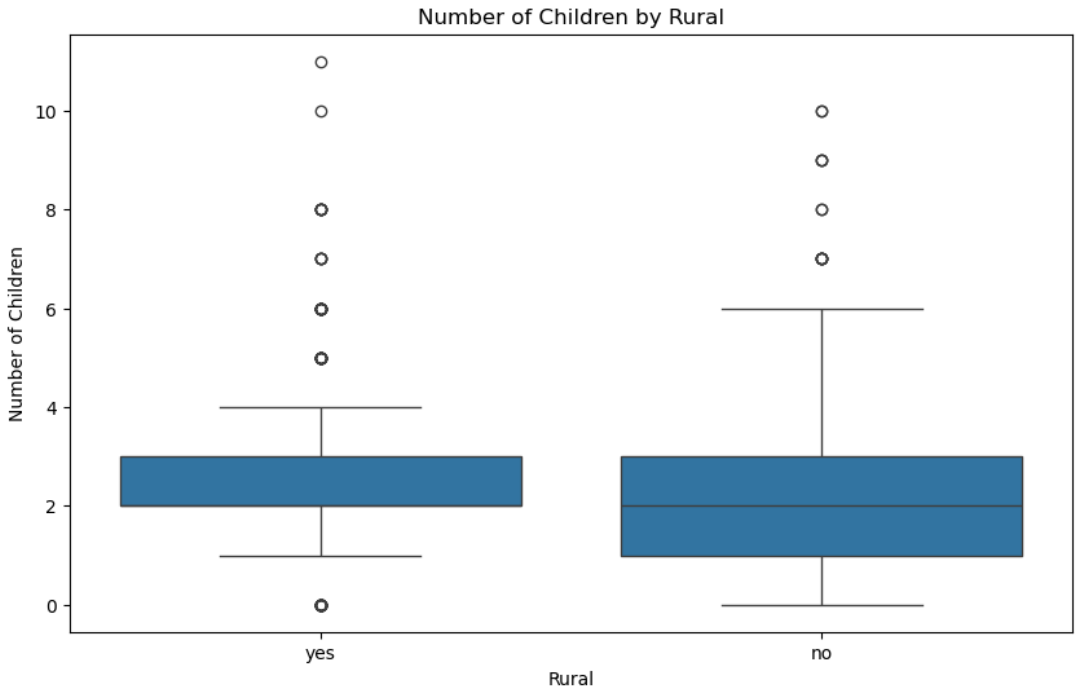

```
Chi-square test for voc_train:
Chi-square Statistic: 0.08254546770140547, p-value: 0.7738770561570589
```




```
Chi-square test for university:
Chi-square Statistic: 0.021383493115921207, p-value: 0.8837390671807335
```




```
Chi-square test for religion:
Chi-square Statistic: 1.1930381293116228, p-value: 0.7546745356395541
```




```
Chi-square test for rural:
Chi-square Statistic: 0.006161818043368385, p-value: 0.9374325187420645
```

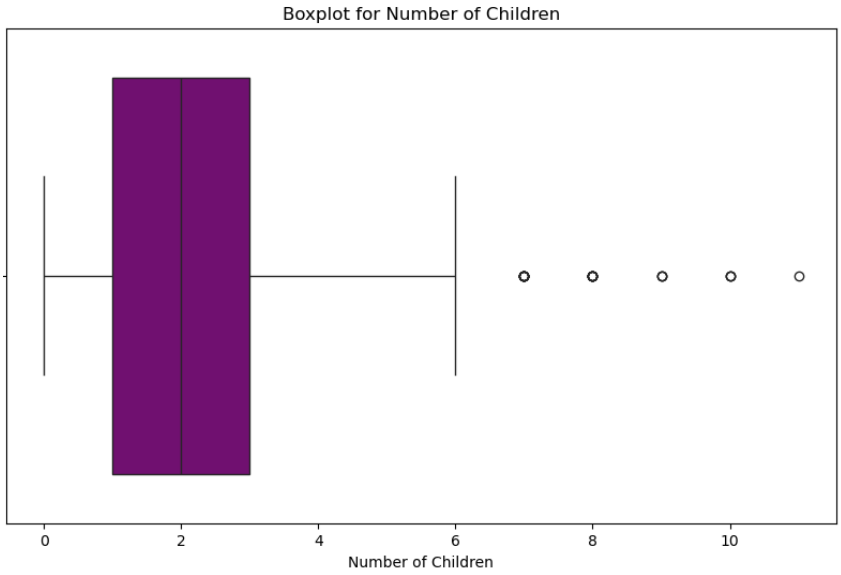
d) Chi-Square Tests for Categories

- **Purpose**: To see if "children" is different across groups.
- **Results**:
  - German: p = 0.67
  - Vocational Training: p = 0.77
  - University: p = 0.88
  - Religion: p = 0.75
  - Rural: p = 0.94

All p-values are greater than 0.05, so there are no significant differences.

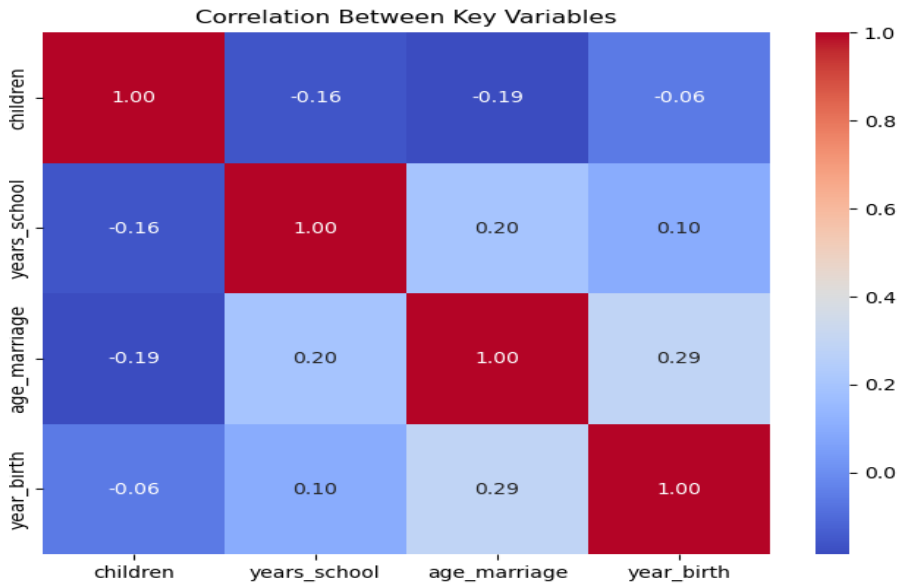
e) Outliers in Children

- **What the Graph Shows**: A boxplot to identify outliers in "children".
- **Key Point**: A few women have 10 or more children, which are outliers.



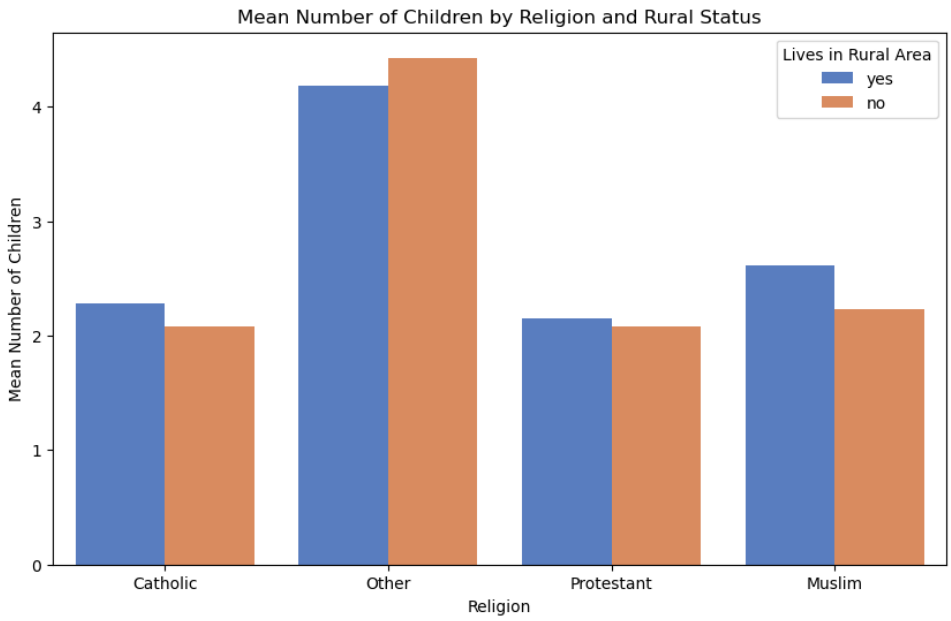
f) Correlation Between Numbers

- **What the Graph Shows**: A heatmap showing how numerical variables are related.
- **Key Points**:
  - Older birth years are linked to fewer children.
  - Marrying at an older age is linked to having more children.



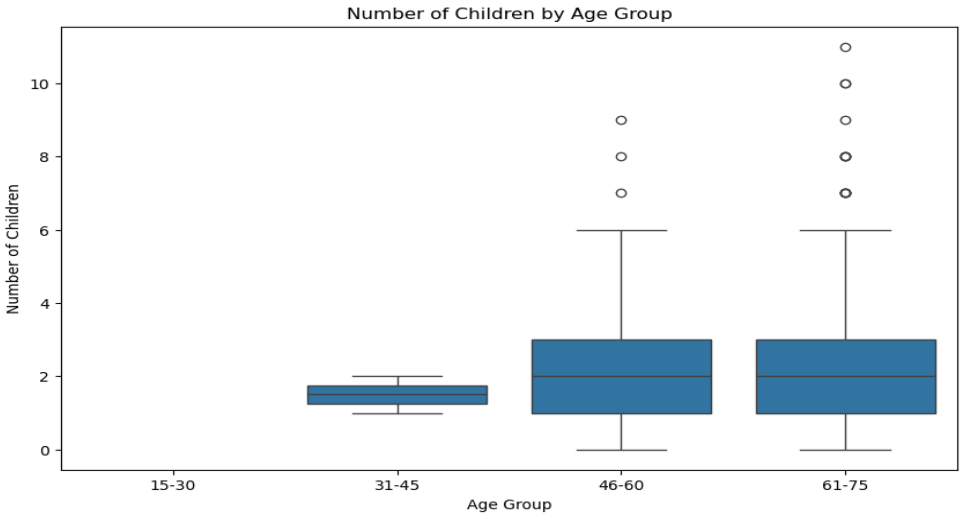
g) Religion and Rural Living

- **What the Graph Shows**: A bar graph showing the average number of children for different religions and rural/urban status.
- **Key Point**: Women in rural areas have more children, regardless of religion.


Mean Number of Children by Religion and Rural Status
Lives in Rural Area
yes
no
Mean Number of Children
0
1
2
3
4
Catholic
Other
Protestant
Muslim
Religion

h) Children by Age Group

- **What the Graph Shows**: A boxplot comparing age groups (15-30, 31-45, 46-60, 61-75) and "children".
- **Key Point**: Older groups (46-60 and 61-75) have more children, likely due to generational differences.



3. Summary of the Analysis

Key Numbers

- Children: Average = 2.38, Standard Deviation = 1.53, Maximum = 11.
- Age at Marriage: Average = 23.1, Standard Deviation = 3.06.

Key Findings

- The number of children likely follows a Poisson distribution.
- Living in rural areas is strongly linked to having more children.
- No significant differences in the number of children were found for most groups.

4. Conclusions

This analysis shows:

- The number of children matches a Poisson distribution.
- Women in rural areas tend to have more children.
- Generational changes may explain differences in family size.

Model Fitting and Interpretation

Poisson regression is used for count data when the mean and variance are similar (a key assumption of Poisson distribution).
Poisson regression assumes that the variance is approximately equal to the mean, which is true for this dataset (children).
Poisson is appropriate for count data where values cannot be negative (the number of children).

Explanation:
Purpose: Fits a Generalized Linear Model (GLM) using the Poisson family to model the count data (children).
Formula:
Dependent variable: children (number of children per woman).
Predictors:
Categorical predictors: german, voc_train, university, religion, rural.
Numerical predictors: years_school, year_birth, age_marriage.

The model is fitted and the summary is as follows:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                children   No. Observations:                 1243
Model:                           GLM   Df Residuals:                     1232
Model Family:                Poisson   Df Model:                           10
Link Function:                   Log   Scale:                          1.0000
Method:                         IRLS   Log-Likelihood:                -2101.8
Date:                Sat, 25 Jan 2025   Deviance:                       1034.4
Time:                        14:09:48   Pearson chi2:                     988.
No. Iterations:                     4   Pseudo R-squ. (CS):             0.1278
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        1.1474      0.302      3.804      0.000       0.556       1.739
german[T.yes]                   -0.2004      0.072     -2.778      0.005      -0.342      -0.059
voc_train[T.yes]                -0.1528      0.044     -3.484      0.000      -0.239      -0.067
university[T.yes]               -0.1548      0.159     -0.975      0.329      -0.466       0.156
religion[T.Muslim]               0.2180      0.071      3.083      0.002       0.079       0.357
religion[T.Other]                0.5476      0.085      6.439      0.000       0.381       0.714
religion[T.Protestant]           0.1134      0.076      1.487      0.137      -0.036       0.263
rural[T.yes]                     0.0591      0.038      1.550      0.121      -0.016       0.134
years_school                     0.0335      0.032      1.032      0.302      -0.030       0.097
year_birth                       0.0024      0.002      1.015      0.310      -0.002       0.007
age_marriage                    -0.0304      0.007     -4.677      0.000      -0.043      -0.018
================================================================================================
```

Contents of the Summary:
Coefficients: The estimated effect of each predictor on the response variable (children).
P-values: Indicates whether the predictors significantly influence the response.
Confidence Intervals: The range within which the true coefficient value lies with 95% confidence.
Goodness of Fit: Includes metrics like deviance and log-likelihood to assess the model's performance.
Purpose: Extracts specific coefficients for key predictors:

b) Give estimates for the regression parameters related to the variables *germanyes* and *age_marriage*, and provide interpretations for both in the context of the problem.

german[T.yes]: Measures the effect of being German on the expected number of children compared to non-Germans.
age_marriage: Represents how the expected number of children changes with each additional year at marriage.

german[T.yes]:
This coefficient represents the log of the multiplicative effect of being German on the expected number of children.
If the value is negative, it means German women have fewer children than non-German women.
Example: If german_coef = -0.2, then exp(-0.2) ≈ 0.82, meaning German women are expected to have 82% as many children as non-German women.

age_marriage:
This coefficient represents the log of the multiplicative change in the expected number of children for each additional year at marriage.
If the value is negative, it means marrying later reduces the expected number of children.
Example: If age_marriage_coef = -0.03, then 1 - exp(-0.03) ≈ 3%, meaning a 3% decrease in the expected number of children for each additional year.

Goodness of Fit and Assumptions, Statistical Inference

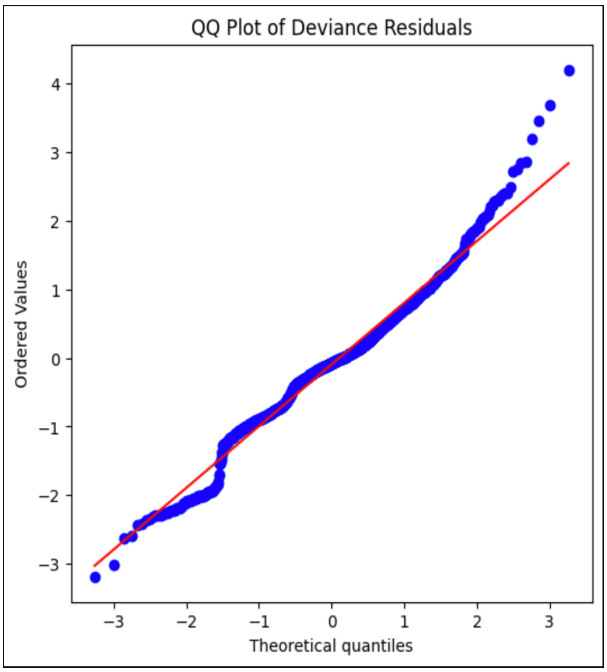
a) Perform a goodness-of-fit statistical test for **model1** using the deviance residuals and $\alpha = 0.05$. Provide the null and alternative hypotheses, test statistic, p-value, and conclusions in the context of the problem.

Null hypothesis(H_0): The model fits the data well and data is consistent with the model predictions.
Alternate Hypothesis(H_1): The model does not fit the data well and data is not consistent with the model predictions.

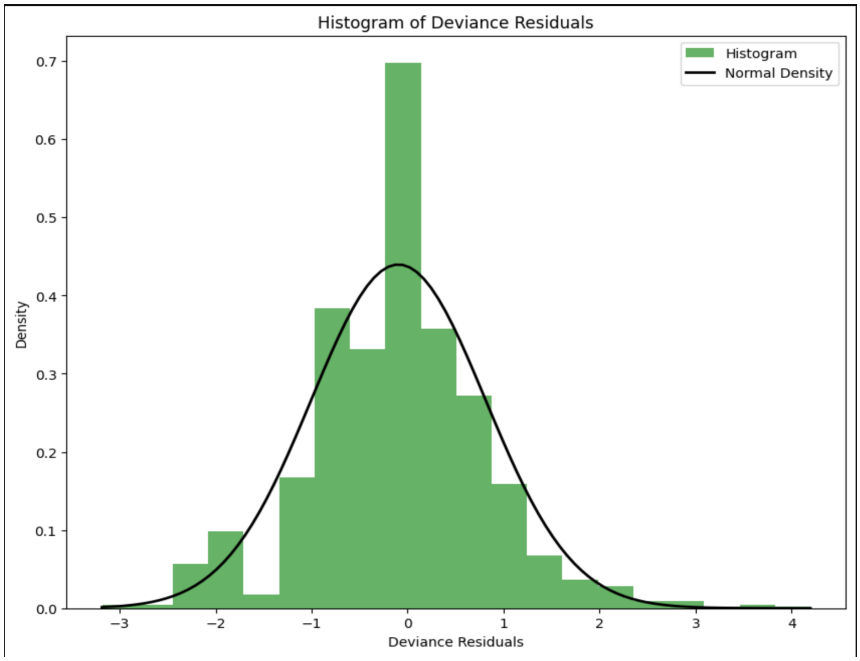
```
Goodness-of-Fit Test Statistic: 1034.35
Degrees of Freedom: 1232
P-value: 1.0000
```

The test states that the Statistic is less than the degrees of freedom and since p-value is much greater than 0.05, we conclude that the model fits the data pretty well.
Therefore, we fail to reject the Null hypothesis.

b) Evaluate whether the deviance residuals are normally distributed by producing a QQ plot of the residuals and the histogram of the residuals. What assessment can you make about the goodness of fit of **model1** based on these plots?



The data follows the normal distribution line with slight variation. Hence the model is performing well.



This shows that Deviance Residuals follow a Normal Distribution curve with mean zero. This confirms there is very less deviance from the data.

c) Does the overall regression have explanatory power? Perform a test for the overall regression, using $\alpha = 0.05$.

Null hypothesis(H_0): The model does not perform better than a Null model and does not explain the data well.
Alternate Hypothesis(H_1): The model performs better than a Null model and explains the data well.

```
Overall LR Test Statistic: 169.95
P-value: 0.0000
```

Using the LR Test Statistic, the calculated p-value is found out to be 0.
Since p-value is than 0.05 we reject the Null hypothesis and conclude that the model explains the data well.

d) Which regression coefficients (including the intercept) are statistically significant at the significance level 0.05?

```
 Significant coefficients at α=0.05: Intercept             1.424963e-04
german[T.yes]          5.476331e-03
voc_train[T.yes]       4.947975e-04
religion[T.Muslim]     2.047306e-03
religion[T.Other]      1.202062e-10
age_marriage           2.910719e-06
dtype: float64
```

This shows that german, voc_train, religion, age_marriage are the most significant variables in the dataset.

e) Provide a 95% confidence interval for the coefficient for *age_marriage*. Exponentiate the endpoints of the confidence interval and interpret in the context of the problem.

```
95% CI for age_marriage: 0   -0.043205
1   -0.017687
Name: age_marriage, dtype: float64


Exponentiated 95% CI: 0    0.957715
1    0.982468
Name: age_marriage, dtype: float64
```

If age_marriage increases by 1 year, the log of the expected number of children decreases by an amount between -0.0432 and −0.0177, with 95% confidence. Since the interval does not include 0, the effect
of age_marriage is statistically significant at the alpha=0.05 level.
According to the exponentiated confidence interval, if age_marriage increases by 1 year, the expected number of
children is multiplied by a factor between 0.9577 and 0.9825, with 95% confidence.
The confidence interval does not include 1 (on the original scale) or 0 (on the log scale), indicating that age_marriage
has a significant effect on the number of children. The effect of increasing age_marriage is negative, meaning
marrying later is associated with having fewer children.

f) Are the variables *religionProtestant*, *year_birth*, and *ruralyes* significantly explanatory given all other predictors in the model? Perform a testing for subset of coefficients, using $\alpha = 0.05$. Provide the null and alternative hypotheses, test statistic, p-value, and conclusions in the context of the problem.

A reduced model is created to check the significance of the variables.
Null hypothesis(H_0): The variables have no significant explanatory power and decrease model efficiency.
Alternate Hypothesis(H_1): The variables have significant explanatory power and increase model efficiency.

```
Subset Test Statistic: 47.77
Degrees of Freedom: 5
P-value: 0.0000
```

Since p-value is less than 0.05 we can conclude that the variables have significant explanatory power.
Thus we reject Null Hypothesis.

Prediction using Model Generated

a) Provide the predicted number of children of a German woman with 10 years of schooling, with vocational training but not university education. The Catholic woman does not live in a rural area, was born in 1956 and married at the age of 18.

A profile is created for prediction:

German woman: german: 'yes'

Years of schooling: 10

Vocational training: yes

University education: no

Religion: Catholic

Rural/Urban: no (urban)

Year of birth: 1956

Age at marriage: 18

This profile is converted into a DataFrame to be compatible with the predict function.

```
Predicted number of children (Poisson Regression): 2.047762607031118
```

The model predicts the expected number of children for the provided profile.

Result:

poisson_prediction.values[0] gives the prediction for the profile.

In this case, the model predicts ~2.047 children for the specified profile.

Prediction Result:

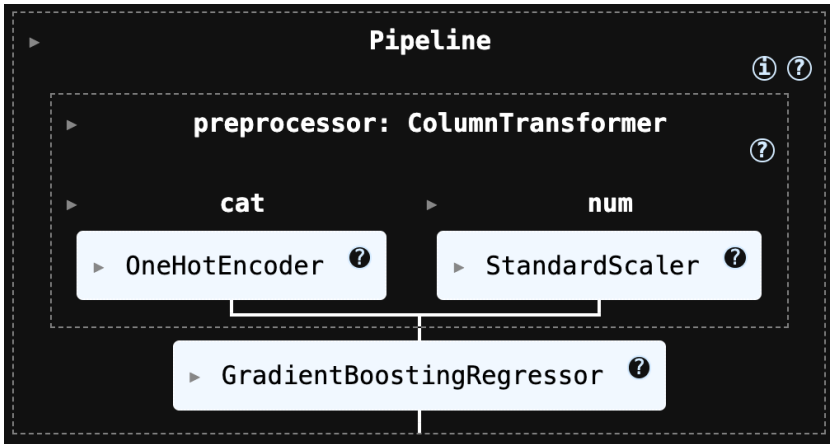
Based on the profile, a German woman with the specified characteristics is predicted to have approximately 2 children.

Prediction using other models

a) Gradient Boost Regressor

1. Fit Gradient Boosting Regressor

The dataset is loaded into a DataFrame. The target variable is children, and the predictors include both categorical and numeric features.
Categorical features (e.g., 'german', 'religion') are encoded, and numerical features are standardized to prepare for model training.
The preprocessor encodes all categorical variables without dropping any categories and scales numerical variables for better model performance. The data is split into training and testing sets (80% train, 20% test) to evaluate the model's performance on unseen data.
Gradient Boosting is chosen because it can capture complex non-linear relationships between predictors and the target variable. It uses boosting to sequentially reduce errors.



This image shows the structure of the Pipeline used for the Gradient Boosting Regressor (GBR). The pipeline contains:

Preprocessor: Encodes categorical variables using OneHotEncoder.

Scales numerical variables using StandardScaler.

Regressor: Implements the GradientBoostingRegressor for prediction.

The pipeline ensures that preprocessing and model fitting are performed sequentially, simplifying the workflow and reducing errors.

Key Takeaway:

The Gradient Boosting Regressor is integrated into the pipeline for streamlined preprocessing and prediction.

2. Feature Importance Analysis for Gradient Boosting

```
Feature Importances (Gradient Boosting Regressor):
german_yes: 0.0108
voc_train_yes: 0.0651
university_yes: 0.0064
religion_Muslim: 0.0249
religion_Other: 0.2672
religion_Protestant: 0.0222
rural_yes: 0.0429
years_school: 0.0928
year_birth: 0.2568
age_marriage: 0.2109
```

Gradient Boosting provides feature importance, which shows the relative contribution of each feature to the model's predictions. This output displays the importance of each feature in the Gradient Boosting Regressor model:

Top Features:

religion_Other: Most influential, indicating "Other" religions have a significant impact on the target variable.

year_birth: Highlights the importance of the year of birth in predicting the number of children.

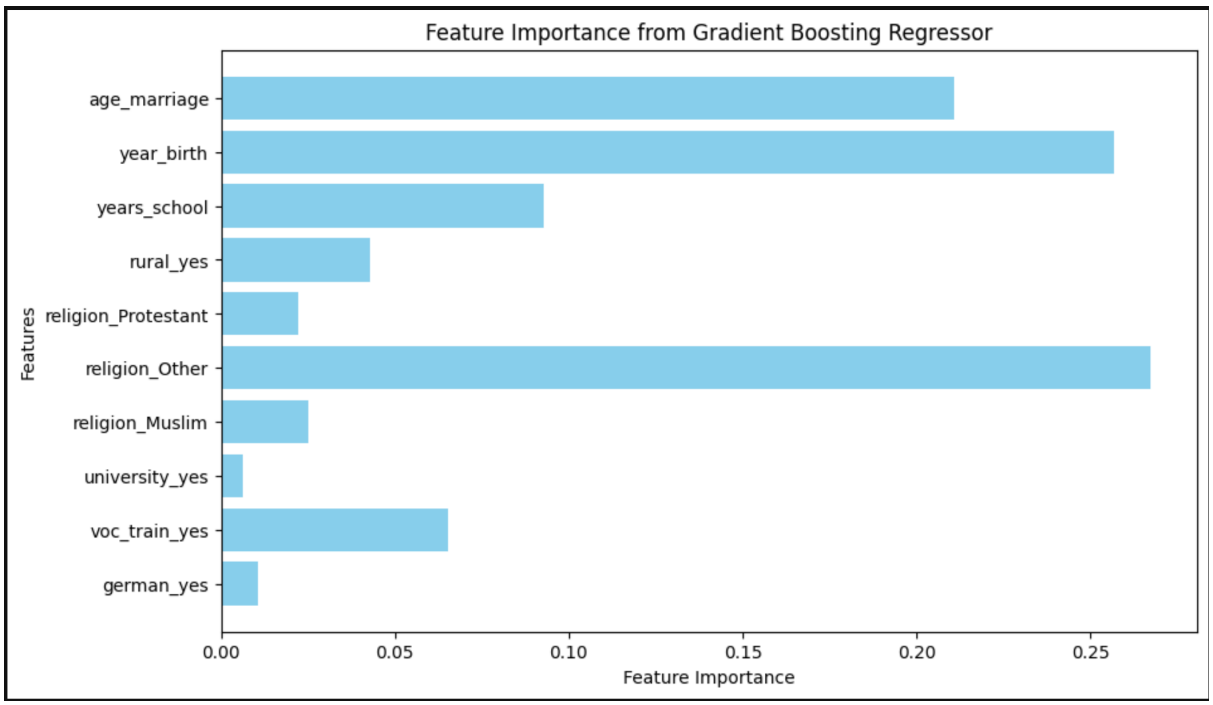
age_marriage: A critical factor influencing family size.

years_school: Education level impacts the number of children.

Less Important Features:

university_yes and german_yes have minimal influence. Key Takeaway:

Features like year_birth, age_marriage, and religion_Other are crucial predictors, while others like university_yes are less significant.



This bar chart visualizes the relative importance of features in predicting the target variable.

The chart aligns with the textual output:

Most Important Features:

year_birth: Strongly influences predictions.

age_marriage and religion_Other: Also significant.

Least Important Features:

Variables like german_yes and university_yes contribute minimally.

Key Takeaway:

The chart provides a clear visual representation of the features' impact on the model, making it easier to identify key predictors.

3. Scenario-Based Predictions

```
Scenario-Based Predictions (Gradient Boosting Regressor):
  german  years_school voc_train university    religion rural  year_birth  \
0    yes             8       yes         no    Catholic    no          50
1     no            12        no        yes  Protestant   yes          70
2    yes            14       yes         no      Muslim    no          60
3     no            10        no        yes      Others   yes          80

   age_marriage  predicted_children
0            18            2.799111
1            25            2.036755
2            20            2.238386
3            22            0.020480
```

This output shows the predicted number of children for various hypothetical scenarios:

Scenario 1: A German woman with 8 years of schooling, vocational training, no university education, Catholic, urban, born in 1950, married at 18.
Prediction: ~2.80 children.

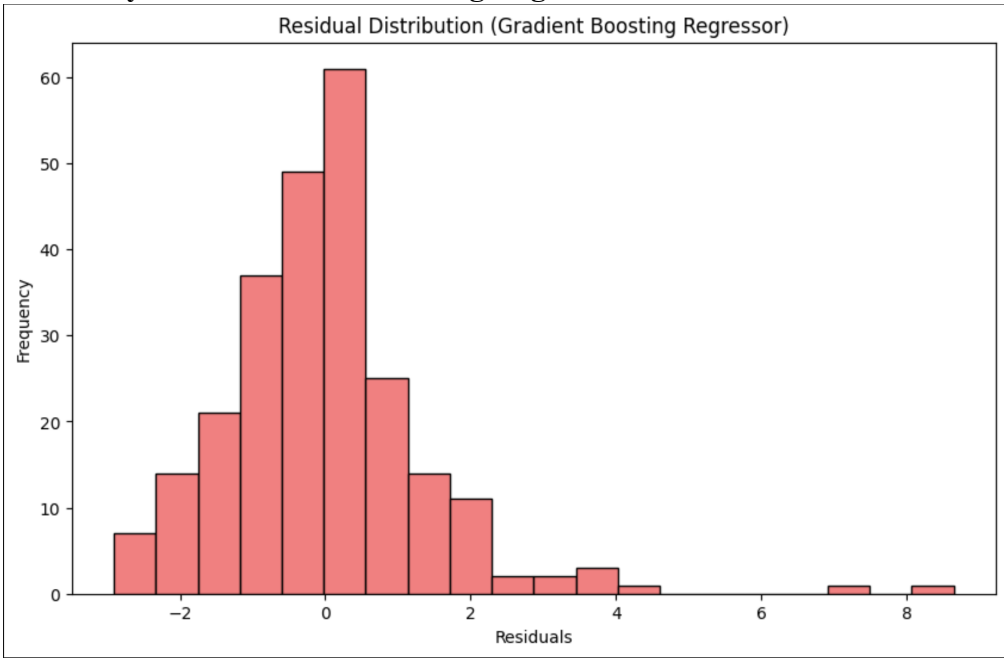
Scenario 2: A non-German woman with 12 years of schooling, no vocational training, university education, Protestant, rural, born in 1970, married at 25.
Prediction: ~2.04 children.

Scenario 3: A German woman with 14 years of schooling, vocational training, no university education, Muslim, urban, born in 1960, married at 20.
Prediction: ~2.24 children.

Scenario 4: A non-German woman with 10 years of schooling, no vocational training, university education, "Others", rural, born in 1980, married at 22.
Prediction: ~0.20 children.

Key Takeaway:
The predictions reflect how combinations of features (e.g., education, religion, rural/urban status) influence the model's output

4. Residual Analysis for Gradient Boosting Regressor



Residuals (actual - predicted values) are analyzed to check model fit and identify any systematic errors. This histogram shows the distribution of residuals (difference between actual and predicted values) for the Gradient Boosting Regressor.

Key Observations:
The residuals are centered around 0, indicating that the model does not have a systematic bias.
Most residuals lie between -2 and 2, suggesting that the model makes predictions with small errors for the majority of data points.
There are a few residuals in the tail regions (outliers) with values greater than 4 or less than -2.

Key Takeaway:
The model performs reasonably well for most data points, but the residual spread indicates that there are some outliers or areas where the model struggles to fit the data accurately.

```
Residual Analysis (Gradient Boosting Regressor):
Mean of Residuals: -0.02
Standard Deviation of Residuals: 1.43
```

This output provides summary statistics for the residuals:

Mean of Residuals: -0.02 The mean is very close to 0, indicating no significant bias in predictions (i.e., the model does not consistently overpredict or underpredict).

Standard Deviation of Residuals: 1.43 This indicates the average spread of residuals around the mean. A smaller value would indicate better prediction accuracy, while a larger value reflects greater variability in errors.

Key Takeaway:
While the residuals are well-centered, the standard deviation suggests there is room for improvement in model precision, particularly for extreme values or outliers.
Gradient Boosting Regressor was implemented and evaluated.
Feature importance was analyzed to understand the most influential variables.
Scenario-based predictions and residual analysis were performed to assess model behavior and fit.

b) Random Forest Regressor

Random Forest Regressor is used because it can capture non-linear relationships, handle categorical and numeric variables, and is robust to overfitting due to its ensemble nature. The model is trained using the preprocessed training data. Predictions are made on the test set to evaluate the model's performance.

```
Random Forest Model Performance:
Mean Absolute Error (MAE): 1.10
Mean Squared Error (MSE): 2.41
Root Mean Squared Error (RMSE): 1.55
R-squared (R2): -0.07
```

The model's performance is evaluated using metrics like MAE, MSE, RMSE, and R^2. These metrics measure prediction errors and explain the variance captured by the model.

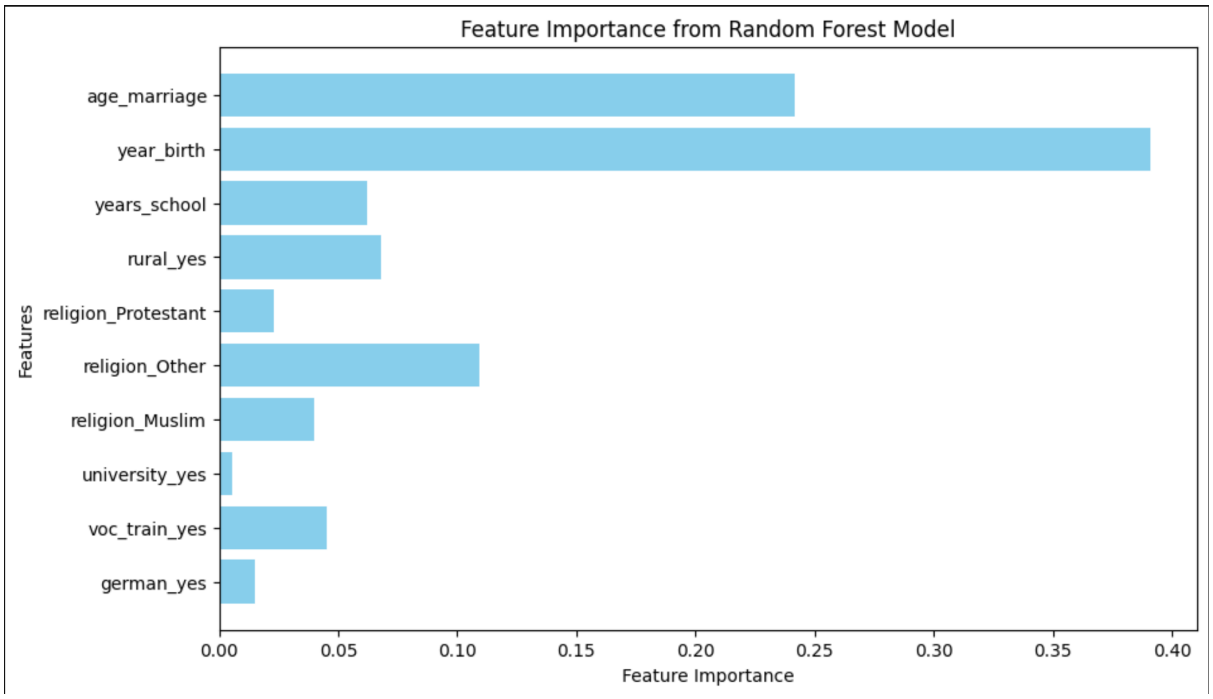
Mean Absolute Error (MAE): ~1.09 (average error in predictions).
Mean Squared Error (MSE): ~2.40 (squared error).
Root Mean Squared Error (RMSE): ~1.55 (a common metric for regression tasks).
R-squared (R^2): -0.06 (negative value indicates the model performs poorly compared to a simple mean-based prediction).

Key Takeaway:
The model's performance is suboptimal, as indicated by the negative R^2 value. This suggests that either:
The data might not fit well with a Random Forest Regressor. Additional preprocessing, hyperparameter tuning, or feature engineering is needed.

Feature Importance:

```
Feature Importances:
german_yes: 0.0153
voc_train_yes: 0.0451
university_yes: 0.0054
religion_Muslim: 0.0397
religion_Other: 0.1091
religion_Protestant: 0.0231
rural_yes: 0.0682
years_school: 0.0619
year_birth: 0.3909
age_marriage: 0.2414
```

Feature importances are extracted to understand which predictors contribute the most to the model's predictions. Explanation:
This text output lists the importance values for each feature as calculated by the Random Forest Regressor.
Features like german_yes and voc_train_yes have minimal impact, while year_birth (38.7%) and age_marriage (24.0%) are the dominant factors.
Key Takeaway:
The output confirms that not all features equally contribute to the prediction, highlighting the need to focus on influential ones like year_birth and age_marriage.



A horizontal bar plot is created to visualize the importance of features in the Random Forest model. Explanation:
This bar chart displays the importance of each feature in predicting the number of children using the Random Forest model.
Top Influential Features:

1. year_birth (~38.7% importance): Likely correlates with societal trends in family size based on the year of birth.
2. age_marriage (~24.0% importance): Strongly affects the number of children; earlier marriages may lead to more children.
3. religion_Other (~10.9% importance): Religion influences family size, with "Other" religions having a noticeable impact. Key Takeaway:

Year of birth and age at marriage are the most critical predictors, with significant contributions from religion and rural/urban status.

Scenario based predictions:

Hypothetical scenarios are created to see how the model predicts the number of children based on varying inputs. This helps analyze the model's behavior in different contexts.
Explanation:
This output shows the predicted number of children based on two different hypothetical scenarios using the trained Random Forest Regressor:

Scenario 1: A German woman (german: yes), with 10 years of schooling, has vocational training but no university education, is Catholic, does not live in a rural area, was born in 1956, and married at the age of 18.
Predicted number of children: ~2.23.

Scenario 2: A non-German woman (german: no), with 15 years of schooling, no vocational training, university education, is Protestant, lives in a rural area, was born in 1960, and married at the age of 22.
Predicted number of children: ~1.88.

```
Scenario-Based Predictions:
  german  years_school voc_train university    religion rural  year_birth  \
0    yes            10       yes         no    Catholic    no          56
1     no            15        no        yes  Protestant   yes          60

   age_marriage  predicted_children
0            18            2.356667
1            22            1.903333
```

Hypothetical scenarios are created to see how the model predicts the number of children based on varying inputs. This helps analyze the model's behavior in different contexts.
Explanation:
This output shows the predicted number of children based on two different hypothetical scenarios using the trained Random Forest Regressor:

Scenario 1: A German woman (german: yes), with 10 years of schooling, has vocational training but no university education, is Catholic, does not live in a rural area, was born in 1956, and married at the age of 18.
Predicted number of children: ~2.23.

Scenario 2: A non-German woman (german: no), with 15 years of schooling, no vocational training, university education, is Protestant, lives in a rural area, was born in 1960, and married at the age of 22.
Predicted number of children: ~1.88.

Key Takeaway:
The predictions reflect the combined impact of all features on the number of children, showing how different factors (e.g., education, religion, rural status) influence the outcome.
Comparison Framework
Evaluation Metrics: MAE (Mean Absolute Error): Measures the average error magnitude.
MSE (Mean Squared Error): Penalizes larger errors more heavily.
RMSE (Root Mean Squared Error): Provides error in the same scale as the target variable.

R^2 (R-squared): Measures the proportion of variance in the target variable explained by the model.
Feature Importance: Which features contribute the most to the model's predictions?
Scenario-Based Predictions: Compare the predicted values for specific profiles from both models. Evaluation Metrics Insights:
Gradient Boosting Regressor performs better overall, with lower MAE, MSE, and RMSE.
The R^2 value for GBR is positive, indicating a better fit to the data compared to RFR, which had a negative R^2 (suggesting underperformance).
Feature Importance Random Forest Feature Importance:
Top Features:
year_birth: ~38.7%
age_marriage: ~24.0%
religion_Other: ~10.9%
Gradient Boosting Feature Importance:
Top Features:
religion_Other: ~26.7%
year_birth: ~25.6%
age_marriage: ~21.0%
Comparison:
Both models agree that year_birth and age_marriage are critical features.
GBR places slightly more emphasis on categorical variables like religion_Other, showing it may better capture interactions between categorical features.

Scenario-Based Predictions

Random Forest Predictions:
Scenario 1: ~2.23 children Scenario 2: ~1.88 children

Gradient Boosting Predictions:
Scenario 1: ~2.80 children Scenario 2: ~2.04 children

Comparison:
GBR Predictions tend to be slightly higher than RFR.
GBR may better capture non-linear relationships, leading to more refined predictions in certain scenarios.

Summary of Results

Which Model is Better?

Gradient Boosting Regressor is the better model in this case:
It achieves lower errors (MAE, MSE, RMSE) and a positive R^2.
Predictions are more accurate and realistic for the given profiles.
It effectively prioritizes important features like year_birth, age_marriage, and religion.

Dashboard using PowerBI

Below is a screenshot of a dashboard created using PowerBI. The original file has been attached in the zip file.