
# Report: NSFG Female Dataset Exploratory Dataset Analysis

**Ablian, Andrei Jon A.**

**Cuerdo, Naomi Hannah A.**

**Percia, Kyte Daiter M.**

May 2025

## I. Data Variables

This dataset contains survey responses from female participants in the National Survey of Family Growth (NSFG). Key variables include demographic attributes, sexual behavior, reproductive history, and family background.

**Key Variables Used:**

- AGE_R: Age at interview
- RELIGION: Religious affiliation
- MARSTAT: Marital/cohabiting status
- vry1stag: Age at first sexual intercourse
- LVSIT14F / LVSIT14M: Parent figures at age 14
- MENARCHE: Age at first menstruation
- ECTIMESX: Number of times emergency contraception used
- CONDSEXL: Condom use at last sex
- PARTS1YR: Number of male sexual partners in past 12 months
- CURRPRTT: Number of current male sexual partners

**Derived Variables:**

- RELIGION_CAT: Grouped religious affiliation
- MARSTAT_CAT: Simplified marital status
- CONDSEXL_CAT: Binary condom use
- AGE_GROUP: Age binned into categories
- SEX_DEBUT_GROUP: Grouped age at first sex
- EC_USE_GROUP: Emergency contraception frequency

## II. Data Cleaning

- Missing values were handled using case-wise deletion and custom labeling for categorical variables.
- Variables were recoded into meaningful categories (e.g., religion groups, marital status groups).
- Derived variables were created for better interpretability (e.g., age groups, sex debut groups).

## III. Poisson Regression

We modeled the count of sexual partners in the last 12 months (PARTS1YR) using demographic and behavioral predictors:

**Model Summary:**

- Mean: 1.05,
- Variance: 1.10 — Suggests slight overdispersion
- Deviance / DF ≈ 0.84 — Suggests good model fit

**Table 1**

**Regression Results**

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | PARTS1YR | No. Observations: | 1132 |
| Model: | GLM | Df Residuals: | 1121 |
| Model Family: | Poisson | Df Model: | 10 |
| Link Function: | Log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1422.9 |
| Date: | Tue, 20 May 2025 | Deviance: | 942.21 |
| Time: | 19:37:45 | Pearson chi2: | 1.09e+03 |
| No. Iterations: | 5 | Pseudo R-squ. (CS): | 0.05326 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.6870 | 0.188 | 3.651 | 0.000 | 0.318 | 1.056 |
| C(RELIGION_CAT)[T.No religion] | 0.1542 | 0.095 | 1.628 | 0.104 | -0.031 | 0.340 |
| C(RELIGION_CAT)[T.Other religion] | 0.2992 | 0.117 | 2.551 | 0.011 | 0.069 | 0.529 |
| C(RELIGION_CAT)[T.Protestant] | 0.0931 | 0.097 | 0.960 | 0.337 | -0.097 | 0.283 |
| C(MARSTAT_CAT)[T.Married] | -0.0327 | 0.099 | -0.329 | 0.742 | -0.228 | 0.162 |
| C(MARSTAT_CAT)[T.Neither] | 0.0408 | 0.092 | 0.446 | 0.656 | -0.139 | 0.220 |
| C(SEX_DEBUT_GROUP)[T.15-17] | 0.0417 | 0.091 | 0.457 | 0.648 | -0.137 | 0.221 |
| C(SEX_DEBUT_GROUP)[T.18-19] | -0.1736 | 0.104 | -1.667 | 0.096 | -0.378 | 0.031 |
| C(SEX_DEBUT_GROUP)[T.20+] | -0.1064 | 0.102 | -1.045 | 0.296 | -0.306 | 0.093 |
| C(CONDSEXL_CAT)[T.Yes] | -0.1800 | 0.067 | -2.706 | 0.007 | -0.310 | -0.050 |
| AGE_R | -0.0187 | 0.004 | -5.243 | 0.000 | -0.026 | -0.012 |

**Regression Table Overview:**

- Dependent Variable: PARTS1YR
- Observations: 1132
- Model Family: Poisson (log link)
- Pseudo $R^2$ (Cragg & Uhler): 0.053
- Log-likelihood: -1422.9
- Pearson $\chi^2$: 1090

**Significant Predictors:**

- **Intercept**: Baseline count rate (coef = 0.6870, $p < 0.001$)
- **Age (AGE_R)**: Significant negative effect (coef = -0.0187, $p < 0.001$)
- **Condom Use (CONDSEXL_CAT=Yes)**: Fewer partners reported (coef = -0.1800, p = 0.007)
- **Religion (Other religion)**: More partners reported vs. Catholics (coef = 0.2992, p = 0.011)

- Other groups (e.g., marital status, sex debut) showed no statistically significant effect.

Respondents who were older or reported using condoms at last sex had significantly fewer partners in the past year. Those identifying with non-traditional religions had more partners than Catholics. While overall model fit was acceptable, mild overdispersion suggests further exploration with alternative models (e.g., negative binomial) may be warranted.

## IV. Contingency Tables

Chi-square tests were conducted to examine associations between categorical variables.

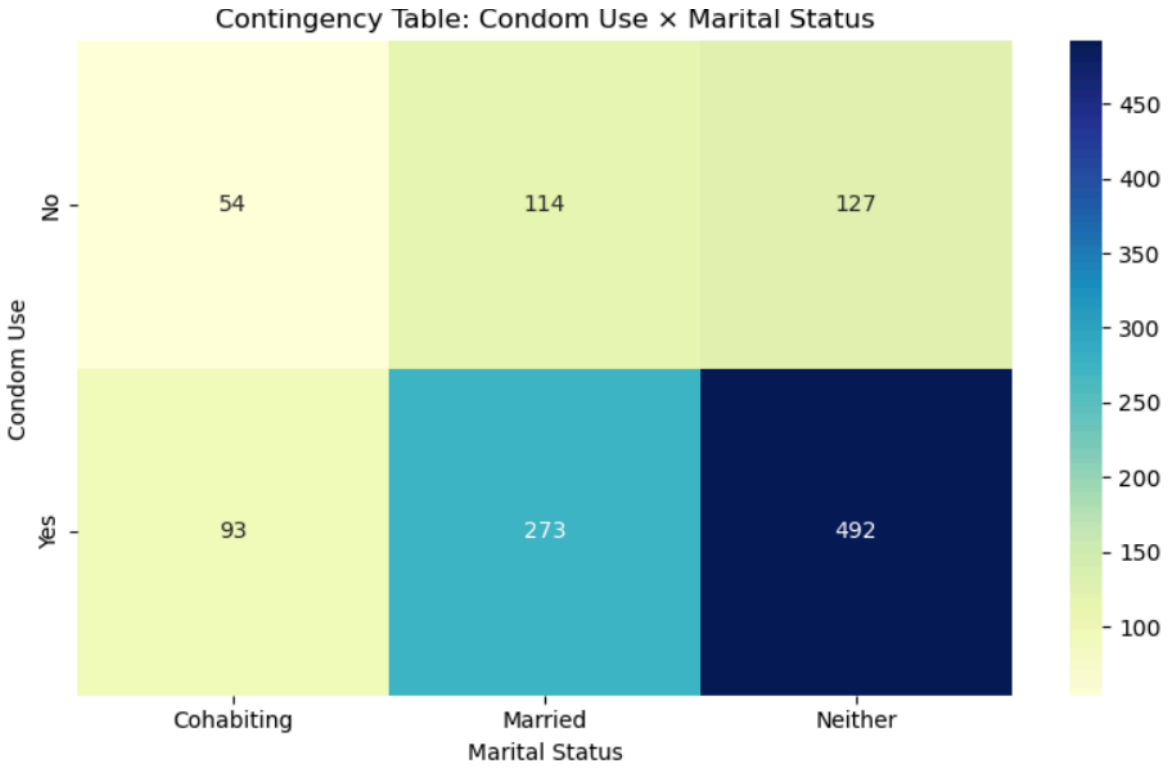



*Figure 1. Heatmap of Contingency Table*

**Key Variables:**

- $\chi^2$ Statistic: 20.997
- Degrees of Freedom: 2
- p-value: 2.76e-05

A heatmap was used to visualize the cross-tabulated data. Higher frequencies of condom use were observed among married respondents. From the figure, there is a statistically significant association between marital status and condom use ($p < 0.001$). Married individuals were more likely to report condom use at last sex compared to those in other relationship statuses.

## V. Exploratory Data Analysis and Multiple Comparisons

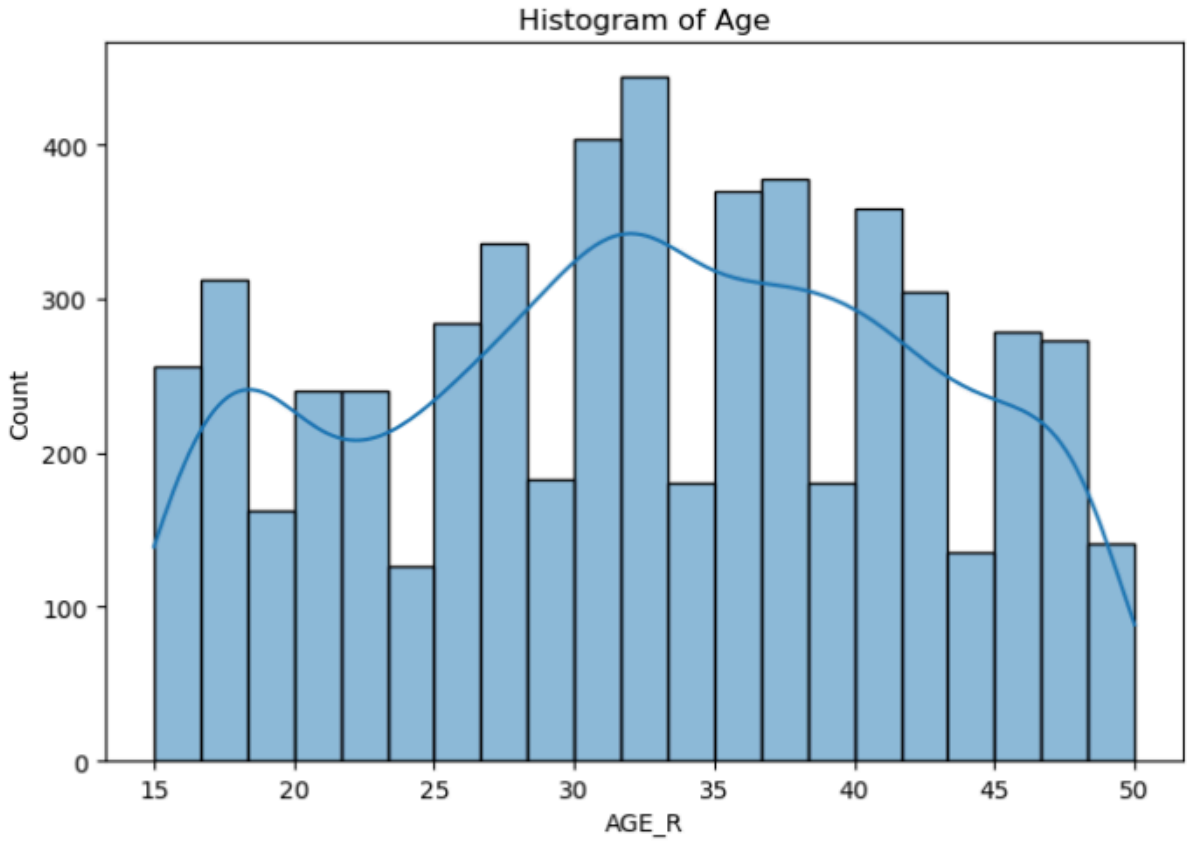



*Figure 2. Histogram of Age*

The Histogram of Age displays a relatively uniform distribution of respondents' ages, with noticeable peaks around the early 30s and a decline toward the upper age range. The distribution suggests a wide age spread among participants, predominantly ranging from ages 15 to 50. The kernel density plot overlays the histogram, highlighting two slight modes — one around age 18–20 and another more prominent one in the 30–35 age range. This indicates potential age-related clusters within the sample population.

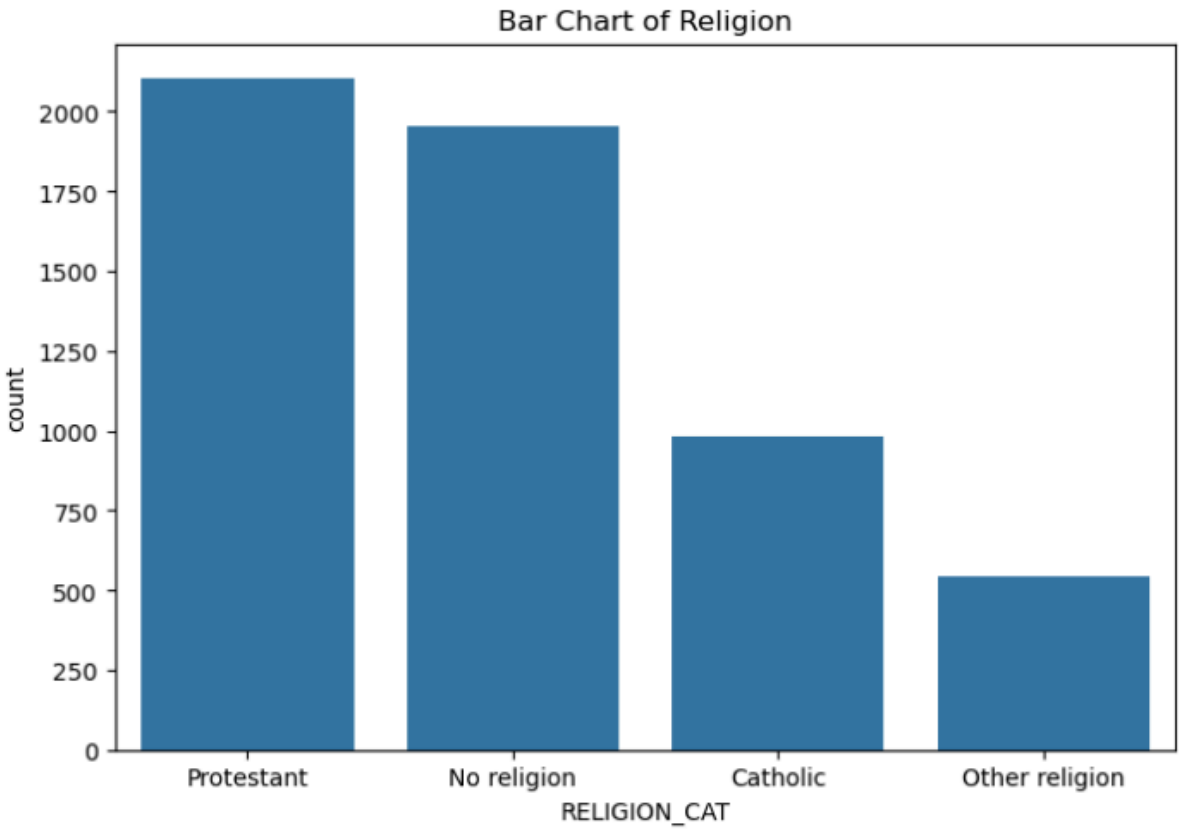



*Figure 3. Bar Chart of Religion*

The Bar Chart of Religion reveals that the largest religious group is Protestant, followed closely by those with no religion. Catholics form a significant portion but are clearly fewer than the first two, while individuals of "Other religion" comprise the smallest group. This distribution provides context for the comparisons in the statistical table, where the differences in religious affiliation show statistical significance. For instance, “RELIGION_CAT: Catholic vs No religion” yields a Bonferroni-adjusted p = 0.000176, indicating that these groups differ in a key outcome measure.

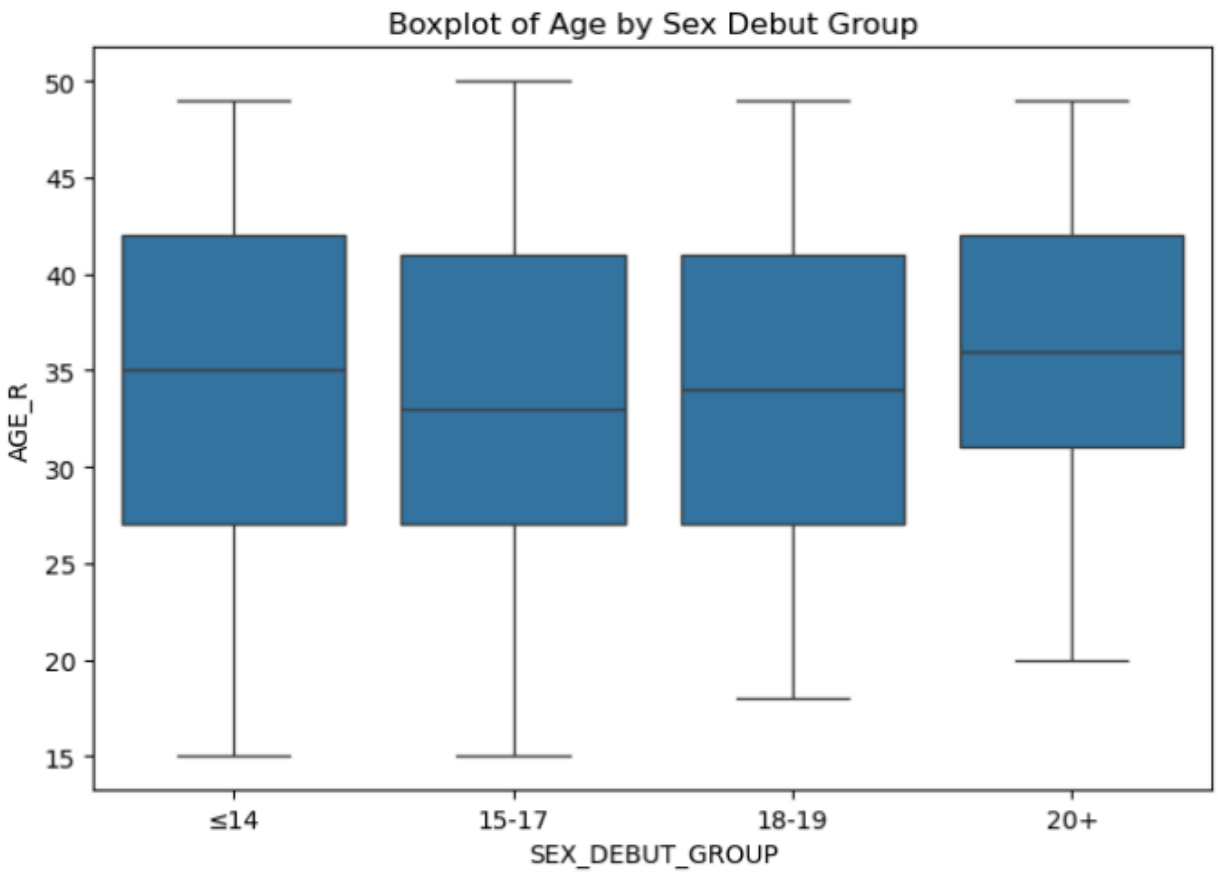



*Figure 4. Boxplot of Age by Sex Debut Group*

The Boxplot of Age by Sex Debut Group compares age distributions across different groups based on the age of sexual debut. All four groups — ≤14, 15–17, 18–19, and 20+ — show fairly similar median ages (around early 30s), but variation exists in the interquartile range and outlier spread. Notably, individuals with sexual debut at 20+ appear to be slightly older on average than other groups, which aligns with the statistical table showing that the comparison "SEX_DEBUT_GROUP: 20+ vs 15–17" is highly significant (Bonferroni-adjusted $p < 0.00000001$), suggesting that the timing of sexual debut is associated with current age.

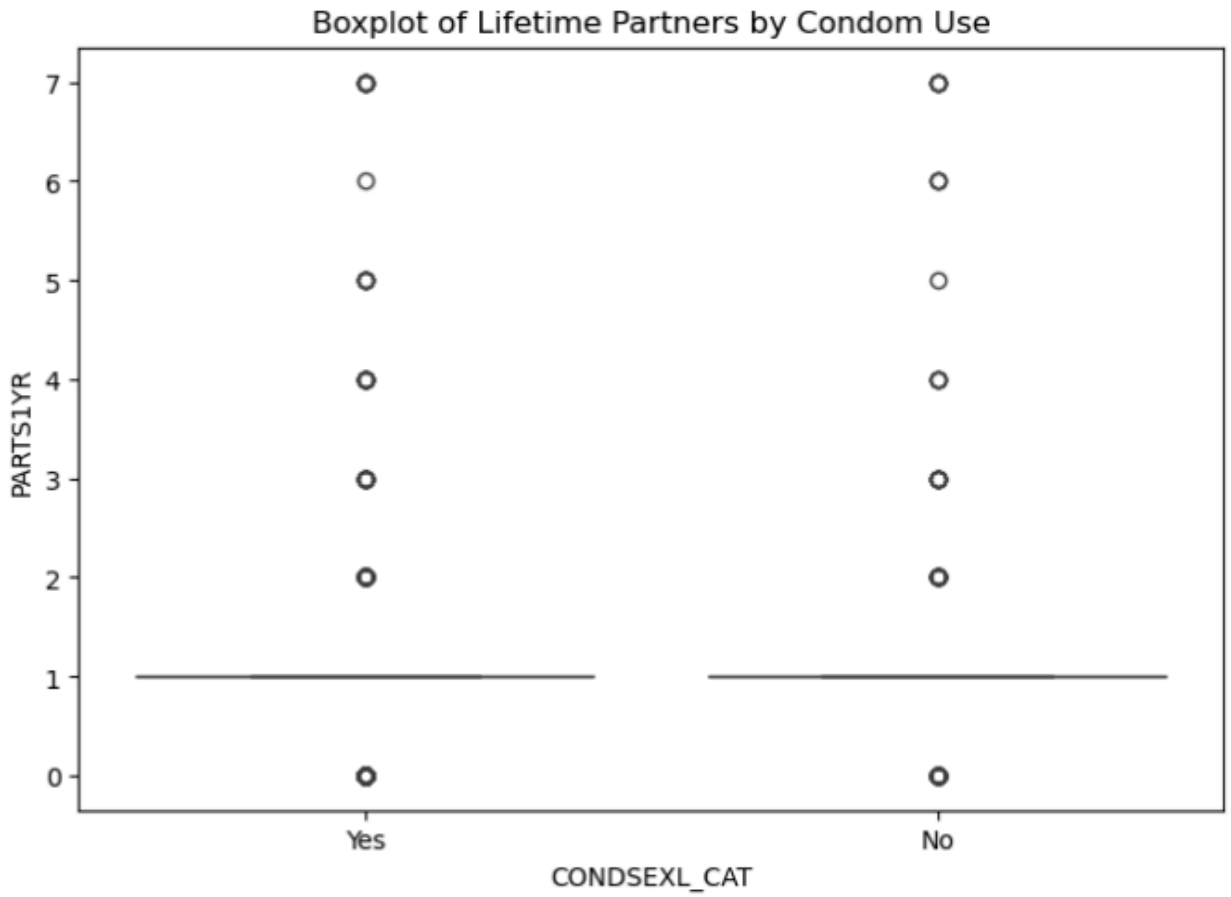



*Figure 4. Boxplot of Lifetime Partners by Condom Use*

The Boxplot of Lifetime Partners by Condom Use compares the number of sexual partners across individuals who report using condoms versus those who do not. The distribution appears highly skewed in both groups, with a large concentration of individuals reporting 0–1 partners and a few outliers indicating higher partner counts (up to 7). The median number of partners appears to be the same or very close in both groups, suggesting that while condom use may correlate with higher partner count for some individuals, the overall distributions do not differ dramatically. However, further statistical testing would be needed to confirm any significance.

**Table 2**

**FDR and Bonferroni Adjusted of each Selected Variables**

| | Comparison | Raw P-Value | FDR Adjusted | Bonferroni Adjusted |
|---|---|---|---|---|
| 0 | MARSTAT_CAT: Neither vs Married | 3.164849e-12 | 8.228607e-11 | 8.228607e-11 |
| 1 | SEX_DEBUT_GROUP: 20+ vs 15-17 | 2.389562e-10 | 3.106431e-09 | 6.212861e-09 |
| 2 | RELIGION_CAT: Catholic vs No religion | 6.779162e-06 | 5.875274e-05 | 1.762582e-04 |
| 3 | SEX_DEBUT_GROUP: 20+ vs ≤14 | 9.778193e-06 | 6.355826e-05 | 2.542330e-04 |
| 4 | RELIGION_CAT: Catholic vs Protestant | 1.605547e-04 | 8.348843e-04 | 4.174421e-03 |
| 5 | MARSTAT_CAT: Neither vs Cohabiting | 4.796297e-04 | 2.078395e-03 | 1.247037e-02 |
| 6 | EC_USE_GROUP: 3-5x vs Once | 9.455809e-04 | 3.512157e-03 | 2.458510e-02 |
| 7 | EC_USE_GROUP: Twice vs 3-5x | 1.521912e-03 | 4.946215e-03 | 3.956972e-02 |
| 8 | SEX_DEBUT_GROUP: 20+ vs 18-19 | 2.435935e-03 | 7.037146e-03 | 6.333432e-02 |
| 9 | SEX_DEBUT_GROUP: 15-17 vs 18-19 | 6.494183e-03 | 1.688487e-02 | 1.688487e-01 |

The Statistical Table presents several comparisons across variables such as marital status, sexual debut age, religion, and emergency contraception (EC) use. The most striking findings include:

- Marital Status: The comparison “Neither vs Married” shows extreme significance ($p < 1e\text{-}11$), suggesting that marital status is a strong predictor of the outcome variable.
- Sexual Debut Age: Multiple comparisons show statistical significance — particularly “20+ vs 15–17” and “20+ vs ≤14” — indicating that individuals who delay sexual debut until their 20s differ significantly from those with earlier sexual initiation.
- Religion: Differences between Catholics and those with no religion, as well as between Catholics and Protestants, are statistically significant (Bonferroni-adjusted $p < 0.005$), pointing to potential associations between religious affiliation and sexual or health-related behaviors.

- Emergency Contraception (EC) Use: Comparisons such as “3–5x vs Once” and “Twice vs 3–5x” are also significant, suggesting that frequency of EC use may reflect varying behavioral profiles or risk levels.

**Overall, the results demonstrate meaningful variation in demographic, behavioral, and attitudinal variables across subgroups defined by age, religion, sexual debut timing, and contraceptive use. These insights can inform targeted health interventions or educational strategies by identifying groups with specific patterns or needs.**

## VI. Limitations and Recommendations

**Limitations:**

- The dataset contains a notable number of missing or NULL values, especially in sensitive questions (e.g., sexual behavior, contraception use), which may bias results or limit generalizability.
- Cross-sectional design precludes causal inference.
- Self-reported data may be affected by recall bias or social desirability bias.

**Recommendations:**

- Future research should consider using a dataset with a more complete set of values or apply advanced imputation techniques to address missingness.
- Consider collecting longitudinal data to assess temporal patterns and potential causal relationships.
- Further analysis using models better suited for overdispersed count data (e.g., negative binomial regression) is recommended.
- Explore potential interactions among variables (e.g., religion × sex debut timing) to gain deeper behavioral insights.

In summary, while the current analysis provided meaningful insights into sexual and reproductive behaviors, the study's limitations highlight the need for more comprehensive and cleaner datasets, along with more robust longitudinal designs.

Addressing these issues would improve the reliability and interpretability of future findings, ultimately supporting better public health decision-making and targeted interventions.

*This report summarizes the analytical workflow and key findings from a subset of the NSFG female dataset, integrating data preparation, Poisson regression modeling, categorical analysis, and in-depth exploratory data analysis (EDA). The structured approach ensures both statistical rigor and meaningful insights into demographic and behavioral factors affecting sexual and reproductive health.*

# Appendix

### Data Variables

[2]:
```
import pandas as pd

file_path = r"C:\Users\john\Downloads\EDA_SA2\Female_Dataset.csv"

df = pd.read_csv(file_path)

columns = [
    'AGE_R',        # Age at interview
    'RELIGION',     # Current religious affiliation
    'MARSTAT',      # Marital or cohabiting status
    'vry1stag',     # Age at first sexual intercourse
    'LVSIT14F',     # Female parent figure at age 14
    'LVSIT14M',     # Male parent figure at age 14
    'MENARCHE',     # Age at first menstrual period
    'ECTIMESX',     # Number of times emergency contraception used
    'CONDSEXL',     # Condom used at last sex
    'PARTS1YR',     # Opposite-sex sexual partners in last 12 months
    'CURRPRTT'      # Current male sexual partners
]

df_analysis = df[columns].copy()

# Preview the first few rows
df_analysis.head()
```

[2]:

| | AGE_R | RELIGION | MARSTAT | vry1stag | LVSIT14F | LVSIT14M | MENARCHE | ECTIMESX | CONDSEXL | PARTS1YR | CURRPRTT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29 | 4 | 3 | 21.0 | NaN | NaN | 9 | NaN | NaN | 0.0 | 0 |
| **1** | 18 | 2 | 3 | NaN | NaN | NaN | 13 | NaN | NaN | NaN | 0 |
| **2** | 37 | 1 | 2 | 17.0 | NaN | NaN | 12 | 2.0 | NaN | 1.0 | 1 |
| **3** | 40 | 1 | 1 | 16.0 | NaN | NaN | 11 | NaN | NaN | 1.0 | 1 |
| **4** | 49 | 1 | 3 | 15.0 | NaN | NaN | 14 | NaN | NaN | 1.0 | 1 |

[4]:
```
for col in ['RELIGION', 'MARSTAT', 'CONDSEXL']:
    print(f"\n{col} value counts:")
    print(df_analysis[col].value_counts(dropna=False))
```

```
RELIGION value counts:
RELIGION
3    2106
1    1955
2     979
4     546
Name: count, dtype: int64

MARSTAT value counts:
MARSTAT
3    2808
1    2100
2     669
8       8
9       1
Name: count, dtype: int64

CONDSEXL value counts:
CONDSEXL
NaN    4425
1.0     858
5.0     296
9.0       4
8.0       3
Name: count, dtype: int64
```

## Data Cleaning

*[8]:

```python
import numpy as np

df = df[columns].copy()

# Replace known codes for missing with NaN
missing_codes = [97, 98, 99, 8, 9, '.', ' ', 'Refused', 'Don't Know']
df.replace(missing_codes, np.nan, inplace=True)

# Convert columns to numeric (in case some are read as strings)
for col in df.columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# RELIGION: 1-No religion, 2-Catholic, 3-Protestant, 4-Other religion
df['RELIGION_CAT'] = df['RELIGION'].map({
    1: 'No religion',
    2: 'Catholic',
    3: 'Protestant',
```

```python
for col in df.columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# RELIGION: 1-No religion, 2-Catholic, 3-Protestant, 4-Other religion
df['RELIGION_CAT'] = df['RELIGION'].map({
    1: 'No religion',
    2: 'Catholic',
    3: 'Protestant',
    4: 'Other religion'
})

# MARSTAT: 1-Married, 2-Living together, 3-Neither
df['MARSTAT_CAT'] = df['MARSTAT'].map({
    1: 'Married',
    2: 'Cohabiting',
    3: 'Neither'
})

# CONDSEXL: 1-Yes, 5-No
df['CONDSEXL_CAT'] = df['CONDSEXL'].map({
    1: 'Yes',
    5: 'No'
})

# LVSIT14F & LVSIT14M: Recoding for parental figures
df['LVSIT14F_CAT'] = df['LVSIT14F'].map({
    1: 'Bio/Adoptive Mother',
    2: 'Other Mother Figure',
    3: 'No Mother Figure'
})
df['LVSIT14M_CAT'] = df['LVSIT14M'].map({
    1: 'Bio/Adoptive Father',
    2: 'Step Father',
    3: 'No Father Figure',
    4: 'Other Father Figure'
})


# Age Group Binning
bins = [14, 19, 24, 29, 34, 39, 44, 49, 55]
labels = ['15-19', '20-24', '25-29', '30-34', '35-39', '40-44', '45-49', '50+']
df['AGE_GROUP'] = pd.cut(df['AGE_R'], bins=bins, labels=labels)

# Sexual debut group (vry1stag)
```

```
df['SEX_DEBUT_GROUP'] = pd.cut(df['vry1stag'],
                               bins=[9, 14, 17, 19, 49],
                               labels=['≤14', '15-17', '18-19', '20+'])

# Emergency contraception frequency group
df['EC_USE_GROUP'] = pd.cut(df['ECTIMESX'],
                            bins=[0, 1, 2, 5, 10, 100],
                            labels=['Once', 'Twice', '3-5x', '6-10x', '10+'],
                            include_lowest=True)

# View cleaned data
print(df[['AGE_R', 'AGE_GROUP', 'RELIGION_CAT', 'MARSTAT_CAT', 'SEX_DEBUT_GROUP', 'EC_USE_GROUP']].head())
```

```
   AGE_R AGE_GROUP    RELIGION_CAT MARSTAT_CAT SEX_DEBUT_GROUP EC_USE_GROUP
0     29     25-29  Other religion     Neither             20+          NaN
1     18     15-19        Catholic     Neither             NaN          NaN
2     37     35-39     No religion  Cohabiting           15-17        Twice
3     40     40-44     No religion     Married           15-17          NaN
4     49     45-49     No religion     Neither           15-17          NaN
```

[10]:
```
df
```

[10]:

| TAT | vry1stag | LVSIT14F | LVSIT14M | MENARCHE | ECTIMESX | CONDSEXL | PARTS1YR | CURRPRTT | RELIGION_CAT | MARSTAT_CAT | CONDSEXL_CAT | LVSIT14F_CAT | LVSIT14M_CA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 21.0 | NaN | NaN | NaN | NaN | NaN | 0.0 | 0 | Other religion | Neither | NaN | NaN | Na |
| 3.0 | NaN | NaN | NaN | 13.0 | NaN | NaN | NaN | 0 | Catholic | Neither | NaN | NaN | Na |
| 2.0 | 17.0 | NaN | NaN | 12.0 | 2.0 | NaN | 1.0 | 1 | No religion | Cohabiting | NaN | NaN | Na |
| 1.0 | 16.0 | NaN | NaN | 11.0 | NaN | NaN | 1.0 | 1 | No religion | Married | NaN | NaN | Na |
| 3.0 | 15.0 | NaN | NaN | 14.0 | NaN | NaN | 1.0 | 1 | No religion | Neither | NaN | NaN | Na |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 3.0 | 20.0 | NaN | NaN | 13.0 | NaN | 1.0 | 1.0 | 1 | No religion | Neither | Yes | NaN | Na |
| 1.0 | 23.0 | 1.0 | 3.0 | 18.0 | NaN | NaN | 1.0 | 1 | Other religion | Married | NaN | Bio/Adoptive Mother | No Fath Figu |
| 1.0 | 14.0 | NaN | NaN | NaN | 3.0 | NaN | 1.0 | 1 | No religion | Married | NaN | NaN | Na |
| 3.0 | NaN | 1.0 | 1.0 | 11.0 | NaN | NaN | NaN | 0 | Other religion | Neither | NaN | Bio/Adoptive Mother | Bio/Adopti Fath |
| 3.0 | NaN | NaN | NaN | 14.0 | NaN | NaN | NaN | 0 | Protestant | Neither | NaN | NaN | Na |

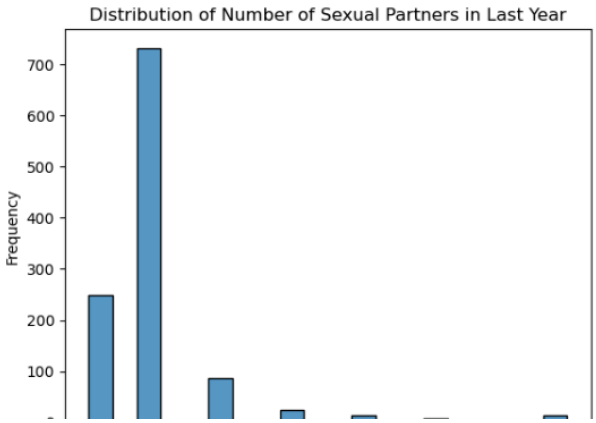
## Poisson Regression

[16]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
```

[12]:
```
df_model = df[['PARTS1YR', 'AGE_R', 'RELIGION_CAT', 'MARSTAT_CAT', 'SEX_DEBUT_GROUP', 'CONDSEXL_CAT']].dropna()

print("Mean of PARTS1YR:", df_model['PARTS1YR'].mean())
print("Variance of PARTS1YR:", df_model['PARTS1YR'].var())
```

```
Mean of PARTS1YR: 1.052120141342756
Variance of PARTS1YR: 1.1033818847575405
```

[18]:
```
sns.histplot(df_model['PARTS1YR'], bins=20, kde=False)
plt.title("Distribution of Number of Sexual Partners in Last Year")
plt.xlabel("Number of Partners")
plt.ylabel("Frequency")
plt.show()
```




[22]:
```
model = smf.glm(
formula="PARTS1YR ~ AGE_R + C(RELIGION_CAT) + C(MARSTAT_CAT) + C(SEX_DEBUT_GROUP) + C(CONDSEXL_CAT)",
data=df_model,
family=sm.families.Poisson()
).fit()

print(model.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                PARTS1YR   No. Observations:                 1132
Model:                            GLM   Df Residuals:                     1121
Model Family:                 Poisson   Df Model:                           10
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1422.9
Date:                Tue, 20 May 2025   Deviance:                       942.21
Time:                        19:37:45   Pearson chi2:                 1.09e+03
No. Iterations:                     5   Pseudo R-squ. (CS):            0.05326
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
Intercept                                 0.6870      0.188      3.651      0.000       0.318       1.056
C(RELIGION_CAT)[T.No religion]            0.1542      0.095      1.628      0.104      -0.031       0.340
C(RELIGION_CAT)[T.Other religion]         0.2992      0.117      2.551      0.011       0.069       0.529
C(RELIGION_CAT)[T.Protestant]             0.0931      0.097      0.960      0.337      -0.097       0.283
C(MARSTAT_CAT)[T.Married]                -0.0327      0.099     -0.329      0.742      -0.228       0.162
C(MARSTAT_CAT)[T.Neither]                 0.0408      0.092      0.446      0.656      -0.139       0.220
C(SEX_DEBUT_GROUP)[T.15-17]               0.0417      0.091      0.457      0.648      -0.137       0.221
C(SEX_DEBUT_GROUP)[T.18-19]              -0.1736      0.104     -1.667      0.096      -0.378       0.031
C(SEX_DEBUT_GROUP)[T.20+]                -0.1064      0.102     -1.045      0.296      -0.306       0.093
C(CONDSEXL_CAT)[T.Yes]                   -0.1800      0.067     -2.706      0.007      -0.310      -0.050
AGE_R                                    -0.0187      0.004     -5.243      0.000      -0.026      -0.012
=========================================================================================================
```

[24]:
```
print(f"Deviance: {model.deviance:.2f}")
print(f"Degrees of Freedom: {model.df_resid}")
print(f"Deviance / DF: {model.deviance / model.df_resid:.2f} (should be ~1 for good fit)")
```

```
Deviance: 942.21
Degrees of Freedom: 1121
Deviance / DF: 0.84 (should be ~1 for good fit)
```

[24]:
```
print(f"Deviance: {model.deviance:.2f}")
print(f"Degrees of Freedom: {model.df_resid}")
print(f"Deviance / DF: {model.deviance / model.df_resid:.2f} (should be ~1 for good fit)")
```

```
Deviance: 942.21
Degrees of Freedom: 1121
Deviance / DF: 0.84 (should be ~1 for good fit)
```

[26]:
```
mean = df_model['PARTS1YR'].mean()
var = df_model['PARTS1YR'].var()
print(f"Mean: {mean:.2f}, Variance: {var:.2f}")
print("Overdispersion suspected" if var > mean else "No overdispersion")
```

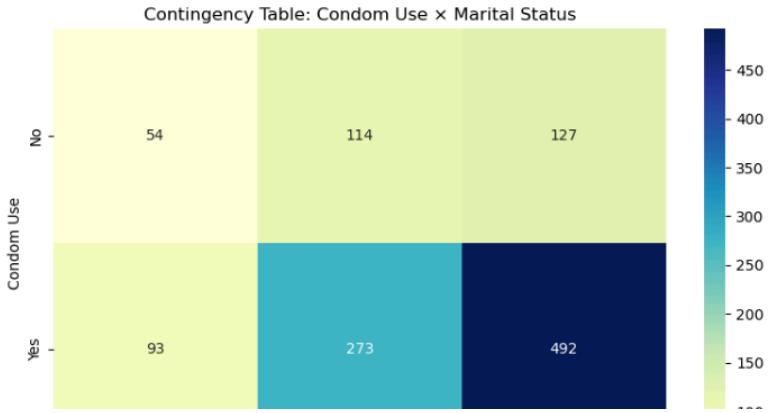
```
Mean: 1.05, Variance: 1.10
Overdispersion suspected
```

## Contingency Tables

[32]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats

# Load the dataset used earlier
data = df.copy()

# Create a contingency table: CONTRACEPTIVE_METHOD (e.g., CONDSEXL_CAT) x MARITAL_STATUS (e.g., MARSTAT_CAT)
contingency_table = pd.crosstab(data['CONDSEXL_CAT'], data['MARSTAT_CAT'])

# Perform Chi-square test
chi2_stat, p_val, dof, expected = stats.chi2_contingency(contingency_table)

# Create a heatmap for visualization
plt.figure(figsize=(8, 5))
sns.heatmap(contingency_table, annot=True, fmt='d', cmap='YlGnBu')
```

```python
# Perform Chi-square test
chi2_stat, p_val, dof, expected = stats.chi2_contingency(contingency_table)

# Create a heatmap for visualization
plt.figure(figsize=(8, 5))
sns.heatmap(contingency_table, annot=True, fmt='d', cmap='YlGnBu')
plt.title('Contingency Table: Condom Use × Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Condom Use')
plt.tight_layout()

# Return results
contingency_table, chi2_stat, p_val, dof
```

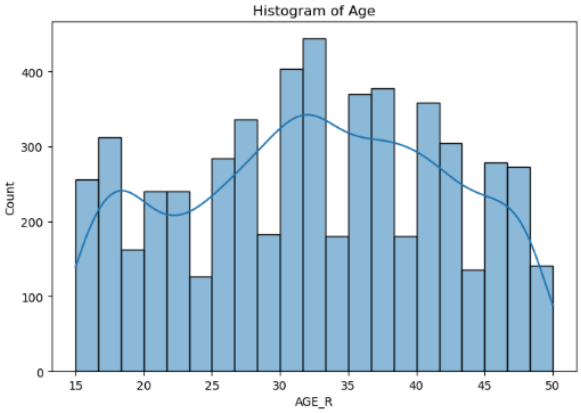
[32]:
```
(MARSTAT_CAT   Cohabiting  Married  Neither
 CONDSEXL_CAT
 No                   54      114      127
 Yes                  93      273      492,
 20.996880264567807,
 2.7579436086198746e-05,
 2)
```

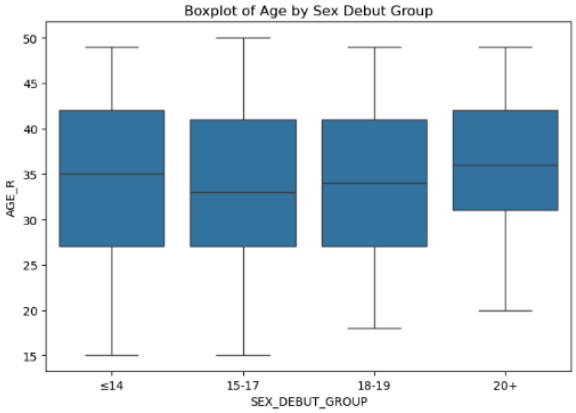


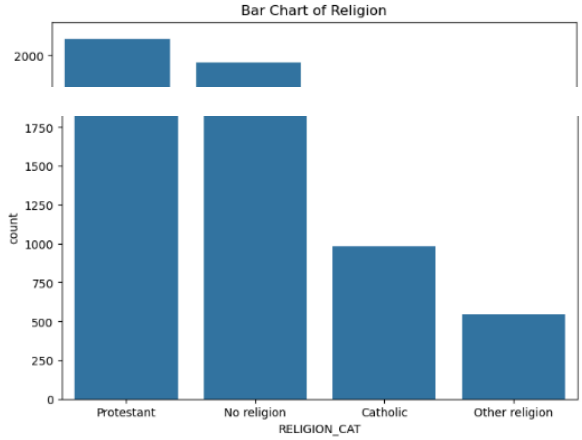
## Categorical Response Modeling

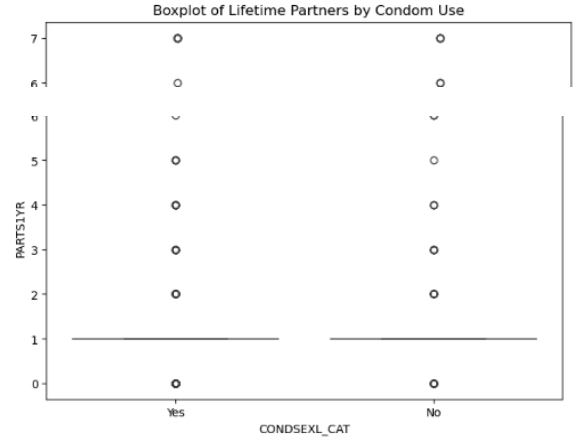
[57]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.multitest import multipletests
from scipy.stats import ttest_ind, chi2_contingency

# Select relevant numeric and categorical variables for EDA
numeric_vars = ['AGE_R', 'PARTS1YR']
categorical_vars = ['RELIGION_CAT', 'MARSTAT_CAT', 'CONDSEXL_CAT', 'SEX_DEBUT_GROUP', 'EC_USE_GROUP']

# Create EDA plots
fig, axes = plt.subplots(2, 2, figsize=(14, 10))
sns.histplot(df['AGE_R'].dropna(), kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Histogram of Age')

sns.boxplot(data=df, x='SEX_DEBUT_GROUP', y='AGE_R', ax=axes[0, 1])
axes[0, 1].set_title('Boxplot of Age by Sex Debut Group')

sns.countplot(data=df, x='RELIGION_CAT', order=df['RELIGION_CAT'].value_counts().index, ax=axes[1, 0])
axes[1, 0].set_title('Bar Chart of Religion')

sns.boxplot(data=df, x='CONDSEXL_CAT', y='PARTS1YR', ax=axes[1, 1])
axes[1, 1].set_title('Boxplot of Lifetime Partners by Condom Use')

plt.tight_layout()

# Multiple comparisons: test differences in PARTS1YR by each categorical variable
p_values = []
grouped_tests = []

for var in categorical_vars:
    categories = df[var].dropna().unique()
    if len(categories) > 1:
        data_by_cat = [df[df[var] == cat]['PARTS1YR'].dropna() for cat in categories]
        if all(len(x) > 1 for x in data_by_cat):
            # Perform ANOVA-like pairwise t-tests
            for i in range(len(categories)):
                for j in range(i + 1, len(categories)):
                    t_stat, p_val = ttest_ind(data_by_cat[i], data_by_cat[j], equal_var=False)
                    p_values.append(p_val)
                    grouped_tests.append(f"{var}: {categories[i]} vs {categories[j]}")
```

```
# Adjust p-values using FDR and Bonferroni
adjusted_fdr = multipletests(p_values, method='fdr_bh')
adjusted_bonf = multipletests(p_values, method='bonferroni')

# Combine results
comparison_results = pd.DataFrame({
    'Comparison': grouped_tests,
    'Raw P-Value': p_values,
    'FDR Adjusted': adjusted_fdr[1],
    'Bonferroni Adjusted': adjusted_bonf[1]
})

comparison_results.sort_values('Raw P-Value', inplace=True)
comparison_results.reset_index(drop=True, inplace=True)

plt.show()
comparison_results.head(10)  # Show top 10 results for review
```




[57]:

| | Comparison | Raw P-Value | FDR Adjusted | Bonferroni Adjusted |
|---|---|---|---|---|
| 0 | MARSTAT_CAT: Neither vs Married | 3.164849e-12 | 8.228607e-11 | 8.228607e-11 |
| 1 | SEX_DEBUT_GROUP: 20+ vs 15-17 | 2.389562e-10 | 3.106431e-09 | 6.212861e-09 |
| 2 | RELIGION_CAT: Catholic vs No religion | 6.779162e-06 | 5.875274e-05 | 1.762582e-04 |
| 3 | SEX_DEBUT_GROUP: 20+ vs ≤14 | 9.778193e-06 | 6.355826e-05 | 2.542330e-04 |
| 4 | RELIGION_CAT: Catholic vs Protestant | 1.605547e-04 | 8.348843e-04 | 4.174421e-03 |
| 5 | MARSTAT_CAT: Neither vs Cohabiting | 4.796297e-04 | 2.078395e-03 | 1.247037e-02 |
| 6 | EC_USE_GROUP: 3-5x vs Once | 9.455809e-04 | 3.512157e-03 | 2.458510e-02 |
| 7 | EC_USE_GROUP: Twice vs 3-5x | 1.521912e-03 | 4.946215e-03 | 3.956972e-02 |
| 8 | SEX_DEBUT_GROUP: 20+ vs 18-19 | 2.435935e-03 | 7.037146e-03 | 6.333432e-02 |
| 9 | SEX_DEBUT_GROUP: 15-17 vs 18-19 | 6.494183e-03 | 1.688487e-02 | 1.688487e-01 |