A PROJECT REPORT

on

Assessment of Some Factors Contributing to Various Sorts of Domestic Violence Against Women in India Using the NFHS-V dataset

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF SCIENCE

in

Statistics

SUBMITTED BY:

SHEIKH ABDULLAH TAUSEEF	20STB005
AZIZUR RAHMAN	20STB020
AZHAR QUADRI	20STB101
MOHD SADIQ RIZVI	20STB104
MONIS	20STB107
ABDUL MOIZ	20STB119
MOHAMMAD HASSAN ZAIDI	20STB121
MOHAMMAD SHABAN IFITIKHAR	20STB122
ZAKARIYYA SHAHID	20STB123
ASHAR NADEEM	20STB124
ABDUL REHMAN KHAN	20STB125
MOHD HUMZA KHAN	20STB126
ALI RAZA	20STB129

Under the supervision of

Prof. Aquil Ahmed

&

Dr. Aijaz Ahmad Dar

DEPARTMENT OF STATISTICS & OPERATIONS RESEARCH

ALIGARH MUSLIM UNIVERSITY

ALIGARH (202002), U.P., INDIA

2023

A PROJECT REPORT

on

Assessment of Some Factors Contributing to Various Sorts of Domestic Violence Against Women in India Using the NFHS-V dataset

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF SCIENCE

in

Statistics

SUBMITTED BY:

SHEIKH ABDULLAH TAUSEEF	20STB005
AZIZUR RAHMAN	20STB020
AZHAR QUADRI	20STB101
MOHD SADIQ RIZVI	20STB104
MONIS	20STB107
ABDUL MOIZ	20STB119
MOHAMMAD HASSAN ZAIDI	20STB121
MOHAMMAD SHABAN IFITIKHAR	20STB122
ZAKARIYYA SHAHID	20STB123
ASHAR NADEEM	20STB124
ABDUL REHMAN KHAN	20STB125
MOHD HUMZA KHAN	20STB126
ALI RAZA	20STB129

Under the supervision of

Prof. Aquil Ahmed

&

Dr. Aijaz Ahmad Dar

DEPARTMENT OF STATISTICS & OPERATIONS RESEARCH

ALIGARH MUSLIM UNIVERSITY

ALIGARH (202002), U.P., INDIA

2023

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY
ALIGARH (202002), U.P., INDIA

CERTIFICATE

This is to certify that the project titled "**Assessment of Some Factors Contributing to Various Sorts of Domestic Violence Against Women in India Using the NFHS-V dataset**" has been carried out by the following group of students of Bachelor's of Science (Statistics), Final Year, 6th Semester, 2022-23, under my supervision for the course "**STB6S1-Project**".

SUBMITTED BY:

SHEIKH ABDULLAH TAUSEEF	20STB005
AZIZUR RAHMAN	20STB020
AZHAR QUADRI	20STB101
MOHD SADIQ RIZVI	20STB104
MONIS	20STB107
ABDUL MOIZ	20STB119
MOHAMMAD HASSAN ZAIDI	20STB121
MOHAMMAD SHABAN IFITIKHAR	20STB122
ZAKARIYYA SHAHID	20STB123
ASHAR NADEEM	20STB124
ABDUL REHMAN KHAN	20STB125
MOHD HUMZA KHAN	20STB126
ALI RAZA	20STB129

Prof. Aquil Ahmed
(Supervisor)

(Chairman)

Dr. Aijaz Ahmad Dar
(Supervisor)

Dedicated to

Late Sir Syed Ahmed Khan

(The Founder of Aligarh

Muslim University,

Aligarh),

Our Parents & Professors

ACKNOWLEDGEMENT

In the Name of **Almighty**, the Most Gracious, the Most Merciful.

All praise and thanks are due to **Almighty**, the Lord of the Worlds, for His blessings and guidance throughout the completion of this project. We thank Him for His infinite mercy and grace.

We would also like to express our deep gratitude to our supervisors, **Prof. Aquil Ahmed** (Chairman) & **Dr. Aijaz Ahmad Dar** (Assistant Professor), Department of Statistics & Operations Research, Aligarh Muslim University, Aligarh for their constant guidance, valuable insights, and support throughout the project. Their expert advice and encouragement have been invaluable to our success.

We also extend our heartfelt thanks to the Chairman of the Department for providing us with all the necessary resources and facilities to carry out our project work.

We are also thankful to the organisation of Demographic & Health Surveys for providing the NFHS-V dataset, without which it was not possible to carry out this piece of work.

Furthermore, we would like to thank our parents, siblings, and colleagues for their unwavering support and encouragement throughout this journey. Their love and encouragement were the driving forces behind our success.

We would express our sincere appreciation to all those who have contributed to this project and helped us in one way or another. May the **Almighty** reward them abundantly.

CONTENTS

Chapter I:
Introduction

1. Introduction

1.1. Background

Over the last three decades, following global events like the United Nations conferences on population and women, the international community has become increasingly aware of the significance of women's social and health status concerning demographic and health outcomes.

Research conducted in the developing world on the connection between violence against women and reproductive health, contributed to raising awareness about the issue and its adverse health outcomes. Acceptance of gender-based violence as a threat to women's health and human rights was made formal when 189 governments signed onto the Platform for Action of the 1995 United Nations' Beijing World Conference on Women. The platform recognizes that violence against women creates obstacles to achieving national-level objectives of equality, development, and peace and violates women's human rights at the individual level. However, the elaboration of programs and monitoring of changes are challenging due to the lack of data and statistics on violence against women.

The definition of "violence against women" includes any act of gender-based violence that results in or may lead to physical, sexual, or psychological harm or suffering for women. The definition covers all forms of violence against women across the entire lifespan. While some forms of violence tend to be specific to a life-cycle stage, other forms cut across all ages. Women can face violence in various forms, like physical violence, sexual abuse, emotional or psychological abuse, verbal abuse, and specific acts of violence during pregnancy. Violence against women can also occur through limiting their access to food and medical care, carrying out dowry deaths, and honor killings, and forcing them to have sex through rape and/or sexual harassment. The perpetrators of violence against women can be intimate partners, family members, or other men, and the subset of violence by intimate partners is usually referred to as "domestic violence," although the term is not always clearly defined.

The Sustainable Development Goals (SDGs) are a set of 17 global goals adopted by the United Nations in 2015 as a universal call to action to end poverty, protect the planet, and ensure that by 2030 all people enjoy peace and prosperity. One of the SDGs that has a direct relationship with domestic violence is SDG 5: Gender Equality. This goal aims to achieve gender equality

NOTE: The content of this chapter is prepared with the help of various books and websites whose references are given in reference section at the end.

and empower all women and girls. The target is to eliminate all forms of discrimination against women and girls and to ensure their full and effective participation and equal opportunities for leadership at all levels of decision-making in political, economic, and public life.

The DHS program aims to provide comparable data on the demographic and health characteristics of populations in developing countries, including nationally representative information on fertility, family planning, infant and child mortality, reproductive health, child health, and the nutritional status of women and children. As violence against women is a health hazard and a barrier to achieving important demographic and health goals, the DHS program has begun collecting data on the prevalence of domestic and other forms of violence against women within the household. However, as of September 2003, only 11 countries that have implemented DHS surveys have collected nationally representative data on domestic violence, with this report providing a summary of findings on domestic violence for nine of these countries.

The DHS survey is an excellent tool for studying not only the linkages between domestic violence and health and demographic outcomes, but also the context in which violence takes place. The DHS Household Questionnaire collects data on household members' sex, age, education, household headship, relationship to the household head, and access to amenities. The DHS Women's Questionnaire collects data on women's characteristics, including age, marital status, parity, contraceptive use, education, employment, and empowerment status, as well as their husband's education, occupation, and alcohol consumption. This information allows for a description of the household context of violence, the characteristics of women who have experienced spousal abuse (and their abusers), and the identification of risk factors stemming from individual, union, and household-level conditions. Throughout this report, the term "domestic violence" is used interchangeably with "spousal violence" or "intimate partner violence," unless otherwise specified. The terms "spouse" and "intimate partner" include any partners with whom the respondent is living or has lived as if married. Additionally, terms such as "currently married" or "ever-married" include "currently partnered" and "ever-partnered" women.

This report aims to provide an insight into gender-based violence, which has not been closely examined empirically. It presents the prevalence of various forms of violence, as well as the characteristics of women who experience violence, their partners, marriages, and households. Moreover, it explores the relationship between violence and indicators of women's empowerment, demographic outcomes, and women's and children's health and nutrition. The main purpose of this document is to bring attention to the issue of gender-based violence, which is believed to have significant linkages to the physical and psychological health of a considerable number of women and children worldwide.

1.2. Domestic Violence

Violence is a behavior that involves the use of physical force, intimidation, or power to harm someone or something. It can take many forms, including physical, sexual, emotional, psychological, or economic violence. The purpose of violence can vary, from gaining control or power over someone or something, to expressing anger or frustration, or inflicting harm for personal gain. Violence can occur in various settings, including within households, communities, and institutions.

Domestic violence is a specific type of violence that occurs within the context of an intimate relationship, such as marriage, cohabitation, or dating. It is used by one partner to gain and maintain power and control over another partner in an intimate relationship. It is a form of violence that can occur in any relationship, regardless of age, gender, sexual orientation, or socioeconomic status. The abuse can be physical, sexual, emotional, psychological, or economic and can take place in a range of settings, including the home, workplace, or public space.

Domestic violence is a significant public health issue with devastating consequences for the victims and their families. It can result in physical injury, chronic health problems, mental health issues, and even death. Domestic violence can also have a profound impact on children who may witness or be exposed to the abuse. It is important to raise awareness about domestic violence, its impact on individuals and society, and to provide support and resources to those affected by it.

1.3. Types of Domestic Violence

1.3.1. Physical Violence

Physical domestic violence refers to any use of physical force against an intimate partner or family member that causes physical harm, injury, or trauma. This type of violence can include hitting, slapping, kicking, choking, punching, or any other physical attack. Physical domestic violence may also involve the use of weapons or objects to harm the victim.

Physical domestic violence can cause serious physical injuries, such as bruises, broken bones, internal bleeding, or permanent disabilities. It can also lead to emotional and psychological trauma, causing the victim to feel scared, anxious, depressed, or helpless.

Victims of physical domestic violence may be afraid to seek help, due to fear of retaliation from their abuser, cultural or societal norms, financial dependence on the abuser, or feelings of shame or guilt. However, it is important for victims to seek help and support, as well as legal protection through restraining orders, to escape the abusive situation and prevent further harm. Additionally, it is important for society as a whole to recognize physical domestic violence as a serious crime and work to prevent it through education, awareness, and support for victims.

1.3.2. Emotional Violence

Emotional or psychological violence in the context of domestic violence involves the use of tactics such as intimidation, manipulation, isolation, humiliation, and verbal abuse to control and dominate a partner or family member. It is often referred to as "mental abuse" or "emotional abuse" and can have a serious impact on the victim's mental and emotional wellbeing.

Examples of emotional or psychological violence in domestic relationships may include threats, name-calling, gaslighting, constant criticism or belittling, withholding affection or support, controlling behaviour, and isolation from family and friends. The abuser may also use tactics such as guilt-tripping, blaming the victim for their actions, or using children to manipulate the victim.

Emotional or psychological violence is often more difficult to recognize and prove than physical violence, as it may not leave visible marks or bruises. However, it can have serious long-term effects on the victim's mental health, leading to depression, anxiety, low self-esteem, and post-traumatic stress disorder (PTSD). It is important to recognize and address emotional or psychological violence in domestic relationships, as it is just as harmful as physical violence and can escalate over time.

1.3.3. Sexual Violence

Sexual violence refers to any sexual act or behaviour that is perpetrated against someone's will, through the use of coercion, threats, intimidation, physical force, or other means. It encompasses a wide range of behaviours, including but not limited to rape, sexual assault, sexual harassment, and other forms of unwanted sexual contact or advances.

Sexual violence can occur within domestic relationships, as well as in other contexts such as workplace, schools, and communities. It can have long-lasting and devastating effects on the victim's physical, emotional, and psychological well-being, and may result in a range of health problems, including sexually transmitted infections, unwanted pregnancies, and mental health disorders such as depression, anxiety, and post-traumatic stress disorder.

Sexual violence is a pervasive problem that affects millions of people worldwide, particularly women and girls. It is a violation of human rights and a public health issue that requires comprehensive prevention and response strategies, including education, awareness-raising, legal and policy reforms, and access to support and services for survivors.

1.4. Causes of Domestic Violence

Domestic violence is a complex issue that can have various causes. These causes can be broadly categorized into individual, relationship, and societal factors. Understanding these underlying causes is crucial to developing effective strategies to prevent and address domestic violence.

1.4.1. Individual Factors

Individual factors refer to personal characteristics that may contribute to domestic violence. Some of these factors include:

- **Low Self-Esteem:** Individuals with low self-esteem may feel powerless and lack control over their lives. This sense of powerlessness can lead to abusive behaviour as a means of asserting power and control over their partners.

- **Substance Abuse:** Individuals who abuse drugs or alcohol are at a higher risk of perpetrating domestic violence. Substance abuse can impair judgment and increase aggression, making it more likely that an individual will engage in violent behaviour.

- **Mental Health Issues:** Individuals with mental health issues, such as depression or anxiety, may have difficulty regulating their emotions and behaviours. This can lead to explosive outbursts and violent behaviour.

- **Childhood Trauma:** Individuals who have experienced abuse or trauma in childhood may be more likely to engage in abusive behaviour as adults. Childhood trauma can also contribute to issues such as low self-esteem, substance abuse, and mental health problems.

1.4.2. Relationship Factors

Relationship factors refer to dynamics within a relationship that may contribute to domestic violence. Some of these factors include:

- **Power and Control:** Domestic violence often involves one partner seeking to assert power and control over the other. This may involve controlling finances, limiting access to friends and family, or using physical violence to maintain control.

- **Jealousy and Possessiveness:** Individuals who are excessively jealous or possessive may engage in abusive behaviour as a means of controlling their partners.

- **Communication Issues:** Poor communication and conflict resolution skills can contribute to domestic violence. When couples are unable to communicate effectively, conflicts can escalate and become violent.

- **Cycle of Violence:** Domestic violence can become a pattern in relationships, with episodes of violence followed by periods of calm. This cycle can be difficult to break without intervention.

1.4.3. Societal Factors

Societal factors refer to broader social and cultural factors that may contribute to domestic violence. Some of these factors include:

- **Gender Inequality:** Domestic violence is often linked to gender inequality, with men seeking to assert their power and control over women. Patriarchal attitudes that reinforce gender roles and norms can contribute to this power imbalance.

- **Social Norms:** Societal norms that condone or excuse violence can contribute to domestic violence. For example, the belief that it is acceptable for men to use violence to assert their authority can contribute to the perpetration of domestic violence.

- **Economic Factors:** Financial stress can contribute to domestic violence. Individuals who are struggling to make ends meet may feel a sense of powerlessness, which can lead to abusive behaviour.

- **Cultural Factors:** Cultural factors, such as attitudes towards marriage and family, can contribute to domestic violence. In some cultures, marriage is viewed as a contract that cannot be broken, which can make it difficult for individuals to leave abusive relationships.

Domestic violence is a complex issue with multiple causes. Understanding these underlying factors is crucial to developing effective prevention and intervention strategies. By addressing individual, relationship, and societal factors, we can work towards creating a world free from domestic violence.

1.5. Sustainable Development Goals (SDGs)

1.5.1. Introduction

The **Sustainable Development Goals (SDGs)**, also known as the Global Goals, are a set of 17 interconnected objectives that were adopted by the United Nations in 2015 as a universal call to action to end poverty, protect the planet, and ensure peace and prosperity for all people by 2030. The SDGs serve as a shared blueprint for peace and prosperity for people and the planet, both now and into the future.

It is critical that no one is left behind in achieving these goals. The SDGs were formally articulated and adopted in a UNGA Resolution called the 2030 Agenda, known colloquially as **Agenda 2030**. Most targets are to be achieved by 2030, although some have no end date. The SDGs recognize that action in one area will affect outcomes in others and that development must balance social, economic, and environmental sustainability.

The 17 Sustainable Development Goals are an urgent call for action by all developed and developing countries in a global partnership. They acknowledge that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth. Achieving these goals is essential for preserving our planet, tackling climate change, and protecting our oceans and forests.

The SDGs are designed to be universal, applying to all countries regardless of their level of development or economic status. They aim to create a more sustainable future for all by encouraging sustainable practices in areas such as energy, agriculture, and infrastructure, while also promoting social and economic equity. Achieving the SDGs requires collaboration and partnership across a range of stakeholders, including governments, civil society, the private sector, and individuals.

1.5.2. Listing of 17 goals

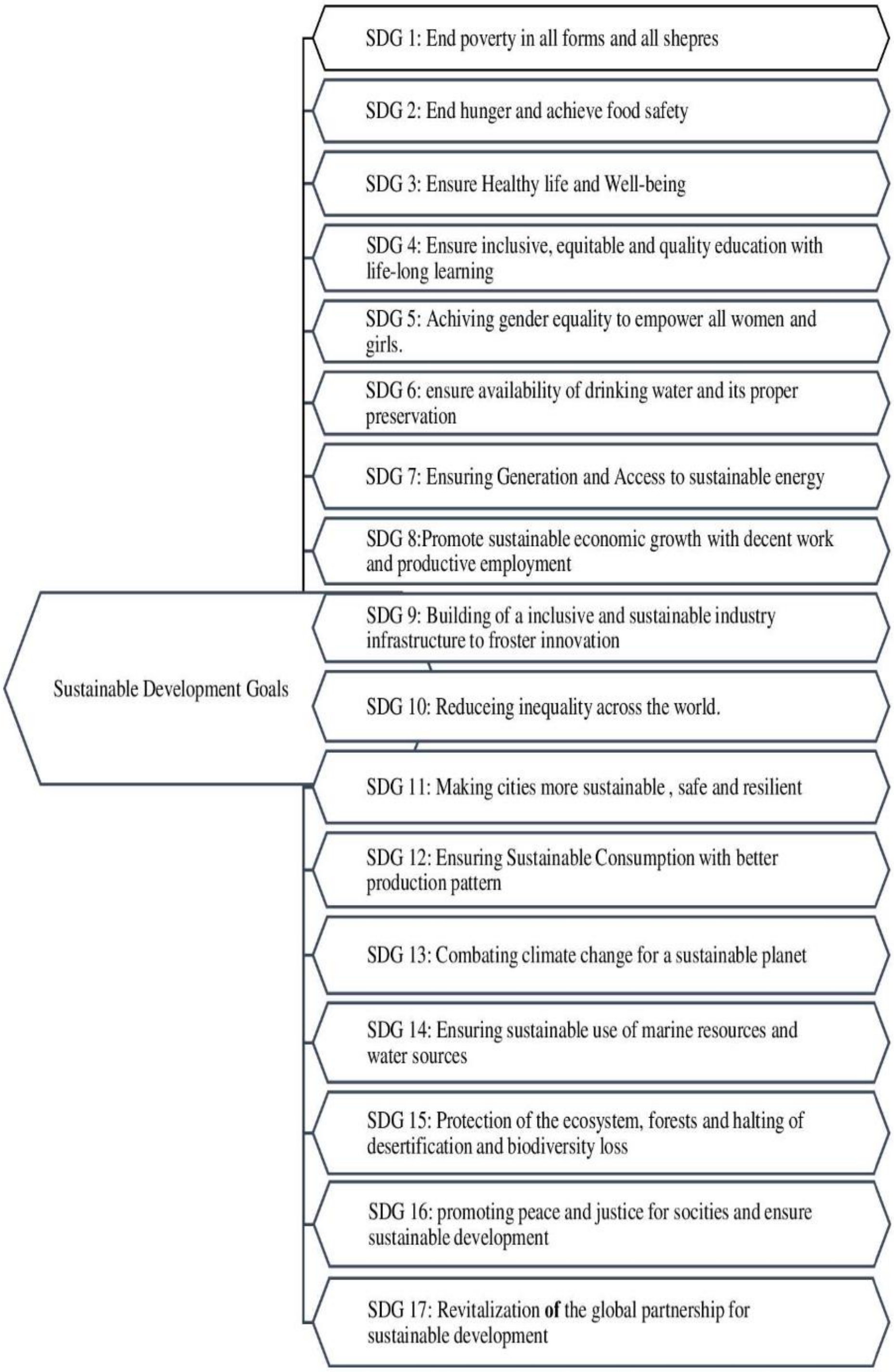


Fig.1 SDG Goals

1.5.3. SDG-5

"***Achieve gender equality and empower all women and girls***"

Gender inequality is a significant factor that contributes to the perpetration of domestic violence. Women and girls are disproportionately affected by domestic violence, which is often fuelled by power imbalances and cultural norms that perpetuate discrimination and oppression. The SDGs recognize the importance of promoting gender equality and the empowerment of women and girls as a means of preventing and eliminating domestic violence.

SDG 5 recognizes the urgent need to eliminate all forms of violence against women and girls, including domestic violence. It sets out several targets and indicators to measure progress in achieving gender equality, including reducing violence against women and girls, promoting women's economic empowerment, and increasing women's participation in leadership and decision-making.

Target 5.2:

"***Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking, sexual exploitation, and other types of exploitation."***

SDG 5 target 5.2 aims to eliminate all forms of violence against women and girls in both the public and private spheres, including trafficking, sexual and other types of exploitation, and harmful practices like child marriage and female genital mutilation. This target recognizes that gender-based violence is a pervasive and systematic human rights violation, affecting women and girls of all ages, socioeconomic backgrounds, and geographic regions, and hindering their ability to fully participate in society and achieve their potential. It also acknowledges that ending gender-based violence requires a comprehensive and multi-sectoral approach that addresses root causes, promotes gender equality and women's empowerment, and strengthens legal and institutional frameworks to prevent and respond to violence. Target 5.2 also recognizes the importance of collecting and using data to monitor progress and inform evidence-based policies and programs. Achieving this target is crucial for the realization of gender equality and the empowerment of all women and girls, as well as for the overall sustainable development agenda.

Indicator 5.2.1:

"Proportion of ever-partnered women and girls aged 15 years and older subjected to physical, sexual, or psychological violence by a current or former intimate partner in the previous 12 months, by form of violence and by age"

Indicator 5.2.1 specifically measures the proportion of ever-partnered women and girls aged 15 years and older subjected to physical, sexual, or psychological violence by a current or former intimate partner in the past 12 months. The indicator aims to provide a quantitative measure of the prevalence of intimate partner violence, which is a significant violation of human rights and a major barrier to gender equality. By tracking progress on this indicator, policymakers can develop evidence-based policies and programs to prevent and respond to intimate partner violence, promote gender equality, and achieve SDG 5.

Chapter II: Research Methodology

2. Research Methodology

2.1. Objective

"In this project, it is aimed to assess various factors contributing to different types of domestic violence against women in India in the light of NFHS-V dataset"

2.2. Dataset

The Demographic and Health Surveys (DHS) Program is a comprehensive and globally recognized program that collects and disseminates data on population, health, and nutrition in low- and middle-income countries. The program is funded by the United States Agency for International Development (USAID) and implemented in partnership with local governments, academic institutions, and international organizations.

The DHS program collects data through nationally representative household surveys, which are conducted using standardized questionnaires and procedures. The surveys cover a wide range of topics, including fertility, family planning, maternal and child health, nutrition, HIV/AIDS, and domestic violence. The DHS program has been in operation for over 30 years, and over 90 countries have participated in the program to date. The program has collected data from over 400 surveys, which have generated over 3,000 datasets. The datasets are publicly available and can be accessed through the DHS program website.

The DHS program is known for its rigorous data collection methods and high-quality data. The surveys use standardized questionnaires that are translated into local languages and pretested to ensure that they are culturally appropriate and easily understood by survey respondents. The surveys are also conducted using standardized procedures, which ensure that the data is comparable across countries and over time. The size of the dataset collected by the DHS program varies depending on the country and the survey round. Generally, the surveys collect data from between 5,000 and 30,000 households, with a total sample size of between 10,000 and 50,000 individuals. The data is collected using a stratified two-stage cluster sampling method, which ensures that the sample is representative of the population at the national and regional levels.

The DHS program is an important resource for researchers, policymakers, and program implementers who are interested in understanding the population and health dynamics of low- and middle-income countries. The data collected by the program is used to inform policies and programs aimed at improving the health and well-being of populations in these countries.

In conclusion, the DHS program is a globally recognized program that collects and disseminates data on population, health, and nutrition in low- and middle-income countries. The program has collected data from over 400 surveys, generating over 3,000 datasets. The data collected is of high quality and is publicly available, making it an important resource for researchers, policymakers, and program implementers. The DHS program plays a critical role in improving our understanding of the health and well-being of populations in low- and middle-income countries.

2.3. Graphical Presentation

2.3.1. Bar Graph

A bar chart, also known as a bar graph, is a graphical representation of data using rectangular bars. The length or height of each bar is proportional to the value it represents, and the bars can be oriented either horizontally or vertically. The x-axis represents the categories being compared, while the y-axis represents the scale or value of the data being measured.

Bar charts are commonly used to compare the values of different categories or groups, as well as to display trends over time. They are often used in business, finance, economics, and other fields to visualize data and make it easier to understand.

Bar charts can be created using various software programs, including Microsoft Excel, Google Sheets, and other data visualization tools. They can be customized with different colors, fonts, labels, and other design elements to make the data more visually appealing and easier to interpret.

2.3.2. Histogram

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of rectangles or bins that are aligned next to each other along an axis. The length of each rectangle or bin corresponds to the frequency or proportion of data that falls within its range.

The bins are typically evenly spaced, although they can be of varying widths if the data is not evenly distributed.

Histograms are commonly used to visualize the shape of a dataset, including the range, central tendency, and variability of the data. They can help identify patterns and outliers in the data, as well as any potential gaps or overlaps. Histograms are also useful for checking the assumptions of statistical models, such as normality and symmetry.

The x-axis of a histogram represents the range of values for the data, and the y-axis represents the frequency or proportion of data that falls within each bin. The bins can be adjusted to be narrower or wider, depending on the granularity of the data and the desired level of detail in the visualization. Histograms are easy to create and interpret and can provide valuable insights into the distribution of the data.

2.3.3. Scatter Plot

A scatter plot (also called a scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables in a set of data. A scatter plot can be used either when one continuous variable is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a parameter exists that is systematically incremented and/or decremented by another, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis, and a scatter plot will illustrate only the degree of correlation (not causation) between two variables. A scatter plot can suggest various kinds of correlations between variables within a certain confidence interval.

2.3.4. Correlation Matrix

A correlation matrix is a table that displays the correlation coefficients for different variables in a dataset. It is used to explore the relationship between variables and identify patterns or trends. Correlation coefficients measure the strength and direction of the linear relationship between two variables. The correlation coefficient ranges from -1 to +1. A coefficient of +1 indicates a perfect positive correlation; a coefficient of 0 indicates no correlation, and a coefficient of -1 indicates a perfect negative correlation.

A correlation matrix is a powerful tool for data analysis as it allows for the identification of strong and weak correlations between variables. It is often used in fields such as finance, economics, and social sciences to identify relationships between different variables, such as the relationship between interest rates and inflation or the relationship between income and education. The correlation matrix can also be used to identify variables that are highly correlated with each other, which can be useful in identifying redundant variables or variables that may be collinear.

Overall, the correlation matrix is a useful tool for exploring relationships between variables in a dataset and can help identify patterns and trends that may be useful in data analysis and decision-making.

2.4. Statistical Tests

2.4.1. Proportion Test

A proportion test is a statistical method used to determine whether the proportion of a particular characteristic or attribute in a sample is significantly different from a known or hypothesized proportion in the population. The proportion test is a hypothesis testing technique that involves comparing the observed proportion in the sample with the hypothesized proportion in the population, using a statistical test such as the z-test or chi-square test.

A proportion test involves comparing two sample proportions to determine if they are statistically different from each other. The null hypothesis in a proportion test is that the two proportions are equal, while the alternative hypothesis is that they are not equal. A test statistic, such as the z-score or chi-square statistic, is used to calculate the probability of obtaining the observed difference in proportions if the null hypothesis is true. If the probability, or p-value, is below a certain threshold, typically 0.05 or 0.01, then the null hypothesis is rejected, and it is concluded that there is a significant difference between the two proportions.

There are several types of proportion tests that can be used depending on the nature of the data and research question:

- **One-sample proportion test:** compares a sample proportion to a known population proportion or a hypothesized proportion
- **Two-sample proportion test:** compares two sample proportions from independent samples
- **Paired proportion test:** compares two sample proportions from paired or matched data, such as before-and-after measurements on the same individuals.
- **Goodness-of-fit proportion test:** tests whether observed proportions in a sample follow an expected distribution or proportion, such as a binomial or normal distribution.

Assumptions of proportion tests

Proportion tests rely on several assumptions, including:

i. The data are binary or dichotomous, meaning there are only two possible outcomes.
ii. The samples are random and independent.
iii. The sample sizes are large enough to assume a normal distribution of the sample proportion, typically at least 10 successes and 10 failures.
iv. The null hypothesis is specified in advance, and the test is appropriately powered to detect a meaningful difference between the proportions, if one exists.

Interpretation of the result

The output of a proportion test typically includes the test statistic, degrees of freedom, p-value, and confidence interval. The p-value indicates the probability of obtaining the observed difference in proportions or a more extreme difference, assuming the null hypothesis is true. A p-value less than the significance level, typically 0.05 or 0.01, indicates that the null hypothesis can be rejected, and that there is evidence of a significant difference between the proportions. The confidence interval provides a range of values within which the true population proportion is likely to fall, based on the sample data and test results.

Limitations of proportion tests

While proportion tests can be useful for comparing proportions and drawing conclusions about differences, they have some limitations:

i. Proportion tests are only as good as the data they are based on. If the data are biased or inaccurate, the results of the test may not be reliable.
ii. Proportion tests do not provide information about the strength or magnitude of the difference between proportions. It only indicates whether the difference is statistically significant or not.
iii. Proportion tests assume a normal distribution of the sample proportion, which may not always be the case in practice. If the sample size is small or the proportion is close to 0 or 1, alternative tests such as Fisher's exact test or a permutation test may be more appropriate.

In conclusion, proportion tests are a powerful statistical tool for comparing proportions and determining whether differences are statistically significant. They have a long history and continue to be used in a wide range of fields, from medical research to quality control. However, as with any statistical method, it is important to use proportion tests appropriately and interpret the results cautiously, taking into account the assumptions and limitations of the test.

2.4.2. Chi-Square Test

The chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, which means that it does not require any assumptions about the distribution of the data being analyzed. This test is widely used in social and health sciences, as well as market research and quality control. The chi-square test is based on the comparison of the observed frequencies of different categories with the expected frequencies, assuming that there is no association between the variables. If the observed frequencies differ significantly from the expected frequencies, then the chi-square test indicates that there is a significant association between the variables. A chi-square test is a valuable tool in research and data analysis, providing insights into the relationships between different categorical variables.

The chi-square test is set up by first defining the null hypothesis and the alternative hypothesis. The null hypothesis states that there is no association between the two categorical variables being compared, while the alternative hypothesis states that there is a significant association between the variables.

Next, a contingency table is created, which shows the observed frequencies of the different categories for each of the variables being analyzed. The contingency table is a two-way table, with the rows representing one variable and the columns representing the other variable. Once the contingency table has been created, the expected frequencies are calculated for each cell of the table, assuming that there is no association between the variables. The expected frequency for each cell is calculated by multiplying the marginal totals for the row and column that the cell belongs to and then dividing by the total sample size.

The chi-square statistic is then calculated by summing the squared differences between the observed and expected frequencies for each cell and dividing by the expected frequency for that cell. The resulting value is compared to a chi-square distribution table with degrees of freedom equal to the product of the number of rows minus one and the number of columns minus one. If the calculated chi-square value is greater than the critical value from the chi-square distribution table, then the null hypothesis is rejected, and it is concluded that there is a significant association between the two variables being compared.

Applications of chi-square tests

There are several applications of chi-square tests that can be used, depending on the nature of the data and the research question:

- **Chi-square goodness-of-fit test:** compares the observed frequencies of a single categorical variable to the expected frequencies from a specified distribution or proportion.
- **Chi-square test of independence:** compares the observed frequencies of two categorical variables to the expected frequencies, assuming no association between them.
- **Chi-square test for homogeneity:** compares the observed frequencies of a categorical variable across multiple populations or groups, assuming no association between the variable and the groups.

Assumptions of chi-square tests

The chi-square test relies on several assumptions, including:

i. The data are categorical, or count data.
ii. The samples are random and independent.
iii. The expected frequencies are greater than 5 for each cell in the contingency table, or 10 for smaller sample sizes.
iv. The null hypothesis is specified in advance, and the test is appropriately powered to detect a meaningful difference between the observed and expected frequencies.

2.4.3. Linear Regression

Regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables. In simpler terms, regression analysis helps to find the relationship between a dependent variable (outcome) and one or more independent variables (predictors). Regression analysis can be used to explore and understand the relationships between variables and to predict the future values of the dependent variable based on the values of the independent variable(s). Regression analysis is often used in many fields, including economics, finance, marketing, biology, psychology, and engineering. It is used to investigate the causal relationship between variables and make predictions based on historical data. There are many types of regression models, including linear regression, logistic regression, polynomial regression, and multiple regression. The choice of model depends on the nature of the data and the research question being addressed.

Linear regression is one of the most widely used regression models. It assumes a linear relationship between the dependent variable and the independent variable(s). The goal of linear regression is to find the best-fit line that explains the relationship between the variables. Multiple regression is used when there are two or more independent variables that affect the dependent variable. The goal of multiple regression is to find the best-fit line that explains the relationship between the variables. Regression analysis is a powerful tool for understanding and predicting the relationships between variables.

A linear model specifies a linear relationship between a dependent variable and independent variables.

$$y = \beta 0 + \beta 1 * X1 + \ldots\ldots\ldots\ldots\ldots + \beta n * Xn + \in$$

where y is the dependent variable, {Xi} are independent variables, {βi} are parameters of the model.

Assumptions of Regression Analysis

i. **Linearity**: The relationship between the dependent variable and the independent variable(s) should be linear.
ii. **Independence**: The observations should be independent of each other.
iii. **Homoscedasticity**: The variance of the errors should be constant across the range of the independent variable(s).
iv. **Normality**: The errors should be normally distributed.
v. **No multicollinearity**: The independent variables should not be highly correlated with each other.

2.4.4. Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical test of the hypothesis that the distribution of the data as a whole deviates from a comparable normal distribution. If the test is non-significant (p>. 05), it tells us that the distribution of the sample is not significantly different from a normal distribution.

The Shapiro–Wilk test tests the null hypothesis that a sample x1, ..., xn came from a normally distributed population.
The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where

- W is the Shapiro-Wilk test statistic,
- n is the sample size,
- $x(i)$ represents the ordered sample values,
- x_i is the i^{th} ordered sample value,
- $\bar{x}$ is the sample mean,
- a_i are the constants specific to the sample size and distribution, and are calculated using statistical software.

The null-hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can't be rejected (e.g., for an alpha level of .05, a data set with a p value of less than .05 rejects the null hypothesis that the data are from a normally distributed population – consequently, a data set with a p value greater than the .05 alpha value fails to reject the null hypothesis that the data is from a normally distributed population).

2.4.5. Variance Inflation Factor (VIF) Test

The variance inflation factor (VIF) is a statistical measure used to assess the level of multicollinearity present in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to inaccurate and unreliable estimates of the regression coefficients. The VIF test determines the degree to which the variance of the estimated regression coefficient is increased due to multicollinearity in the data. A high VIF value indicates a high degree of multicollinearity between the independent variables, which can distort the results of the regression analysis. Generally, a VIF value of 1 indicates no correlation, while a value greater than 5 or 10 indicates a strong correlation and suggests that the independent variable should be removed from the model. The VIF test is commonly used in linear regression analysis, where it is used to identify and address the issue of multicollinearity to improve the accuracy and reliability of the regression model.

2.5. Empirical Analysis of the Odds Ratio

It is a statistical measure used to compare the odds of an event occurring in one group to the odds of the same event occurring in another group. The odds ratio is calculated by dividing the odds of an event occurring in one group by the odds of the same event occurring in another group. If the odds ratio is greater than 1, it indicates that the event is more likely to occur in the first group, while an odds ratio of less than 1 indicates that the event is more likely to occur in the second group. An odds ratio of 1 indicates no difference in the odds of the event occurring in either group.

One limitation of the odds ratio is that it can be difficult to interpret. The odds ratio can be complex, and the results may be difficult for non-experts to understand. In addition, the odds ratio is often misinterpreted as the relative risk, which can lead to erroneous conclusions.

The odds ratio is an important measure of association in observational studies, but it does not provide information about causation. Therefore, it should be interpreted with caution and in conjunction with other statistical measures.

Assumptions of the Odds Ratio

i. Independence of the events being compared: This assumption means that the occurrence of one event does not affect the occurrence of another. If the events are not independent, the odds ratio may be biased.
ii. Absence of confounding variables: Confounding variables are factors that are related to both the exposure and the outcome and can distort the association between the two. To obtain a valid odds ratio, it is important to control for confounding variables through study design or statistical methods.
iii. Adequate sample size: A large enough sample size is needed to ensure that the results are statistically significant and to minimize the effect of chance.

Chapter III: Statistical Analysis

3. Statistical Analysis

3.1. Proportion Test

A proportion test is a statistical method used to determine whether the proportion of a particular characteristic or attribute in a sample is significantly different from a known or hypothesized proportion in the population. The proportion test is a hypothesis testing technique that involves comparing the observed proportion in the sample with the hypothesized proportion in the population, using a statistical test such as the z-test or chi-square test.

3.1.0. Generalized Table

The following general 2x2 contingency table is now subjected to the proportion test in order to compare different classes.

	Violence	**No Violence**	**Total**
C1	n(V,C1)	n(NV,C1)	n(C1)
C2	n(V,C2)	n(NV,C2)	n(C2)
	n(V)	n(NV)	Total

In this part, the proportion test will be used to determine whether there is a significant difference between the proportion of violent crime in the C1 and C2 classes.

The hypothesis will be set as follows:

H_0: $P_{V,C1} = P_{V,C2}$
H_1: $P_{V,C1} \neq P_{V,C2}$

The corresponding test statistic can be written as:

$$Z = \frac{P_{v,C1} - P_{v,C2}}{\sqrt{V(P_{v,C1} - P_{v,C2})}}$$

Where,

$P_{v,c1} = \frac{n(v,c1)}{n(c1)}$ i.e., proportion of violence in C1 class.

$P_{v,c2} = \frac{n(v,C2)}{n(C2)}$ i.e., proportion of violence in C2 class.

R-programming is used to apply the above-mentioned proportion test to all of the classes, and the results are presented in the table below.

The classification of the cases is done based on various criteria/attributes, viz., ***respondent's partner drinks*** (Yes/No), ***respondent's partner is educated*** (Yes/No), ***respondent has children*** (Yes/No), ***respondent is rural*** (Yes/No), ***respondent belongs to a lower caste*** (Yes/No), ***partner is related to respondent prior to marriage*** (Yes/No), ***respondent is working*** (Yes/No), ***can say no to her husband for intercourse*** (Yes/No), ***pregnancy ended naturally*** (Yes/No), ***respondent belongs to the upper region*** (Yes/No), ***respondent is educated*** (Yes/No), *and* ***respondent belongs to BPL*** (Yes/No).

The presence of the attribute is denoted by class C1 and its absence is denoted by class C2 in other words we can denote C1 and C2 as follows

C1: Criterion fulfilled/ presence of attribute
C2: Criterion not fulfilled/ absence of attribute.

Since there are various criteria for classification and it would not be wise to write all the sets of hypotheses separately, we can define them in the following general way.

➢ **Hypothesis set up**

H_o: proportion of violence in C1 = proportion of violence in C2.
H_1: proportion of violence in C1 ≠ proportion of violence in C2.

➢ **Table**

Criterion	$P_{v,c1}$	$P_{v,c2}$	P-value	Decision
Respondent is educated	0.17	0.24	2.2e-16	Reject null hypothesis
Partner is educated	0.18	0.24	2.2e-16	Reject null hypothesis
Partner drinks	0.29	0.17	2.2e-16	Reject null hypothesis
Respondent working	0.18	0.15	4.353e-16	Reject null hypothesis
Respondent belongs to Below Poverty Line	0.22	0.16	2.2e-16	Reject null hypothesis
Place of residence	0.20	0.16	2.2e-16	Reject null hypothesis
No. of Children	0.13	0.19	2.2e-16	Reject null hypothesis
Pregnancy ended naturally	0.20	0.24	0.03285	Reject null hypothesis

Related to husband prior to marriage	0.20	0.16	2.2e-16	Reject null hypothesis
Can you say no to your husband for intercourse	0.16	0.15	0.0275	Reject null hypothesis
Belongs to lower caste	0.20	0.14	2.2e-16	Reject null hypothesis

3.1.1. Violence v/s Respondent's Education Level

- **Hypothesis setup**

Ho: Proportion of violence among those who are educated = proportion of violence among those who are not educated.
H1: Proportion of violence among those who are educated ≠ proportion of violence among those who are not educated.

- **Proportion Test**

Highest educational level * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Highest educational level	No education	11601	3687	15288
	Educated	32812	6893	39705
Total		44413	10580	54993

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who are not educated.
x1 represents the number of victims who faced violence and are not educated.
n1 represents the total cases of not being educated.
P2 represents the proportion of violence among those who are educated.

x2 represents the number of victims who faced violence and are educated.
n2 represents the total cases of being educated.

We will use R for calculating the proportion with the command being:

> prop.test(c(3687,6893),c(15288,39705))

- **Output**

2-sample test for equality of proportions with continuity correction

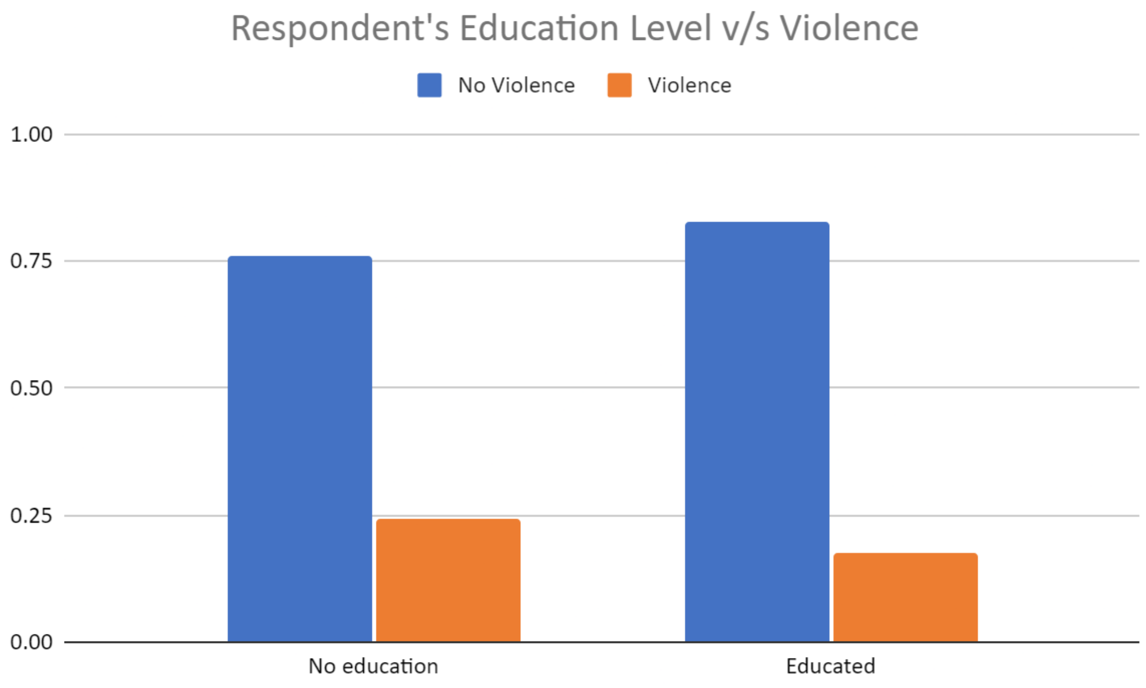
data: c(3687, 6893) out of c(15288, 39705)
X-squared = 323.86, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
0.05978166 0.07534675
sample estimates:
prop 1 prop 2
0.2411695 0.1736053

- **Result**

As the p-value is less than 0.05 the null hypothesis is rejected and we conclude that:

1. Proportion of violence among those who are educated ≠ Proportion of violence among those who are not educated.
2. Respondent's education level has a significant effect on violence.
3. The proportion of violence among those who aren't educated is 0.2411695.
4. The proportion of violence among those who are educated is 0.1736025.

- **Graph**



3.1.2. Violence v/s Respondent's Husband/partner education level

- **Hypothesis setup**

Ho: Proportion of violence among those whose husbands are educated = proportion of violence among those whose husbands are not educated.
H1: Proportion of violence among those whose husbands are educated ≠ proportion of violence among those whose husbands are not educated.

- **Proportion Test**

Husband/partner's education level * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Husband/partner's education level	No education	6999	2276	9275
	Educated	37257	8267	45524
Total		44256	10543	54799

P1 = x1/n1
P2 = x2/n2

Where,
P1 represents the proportion of violence among those whose husbands are not educated.
x1 represents the number of victims who faced violence by uneducated husbands.
n1 represents the total cases of not being educated.
P2 represents the proportion of violence among those whose husbands are educated.
x2 represents the number of victims who faced violence by educated husbands.
n2 represents the total cases of being educated.

We will use R for calculating the proportion with the command being:

```
> prop.test(c(2276,8267),c(9275,45524))
```

- **Output**

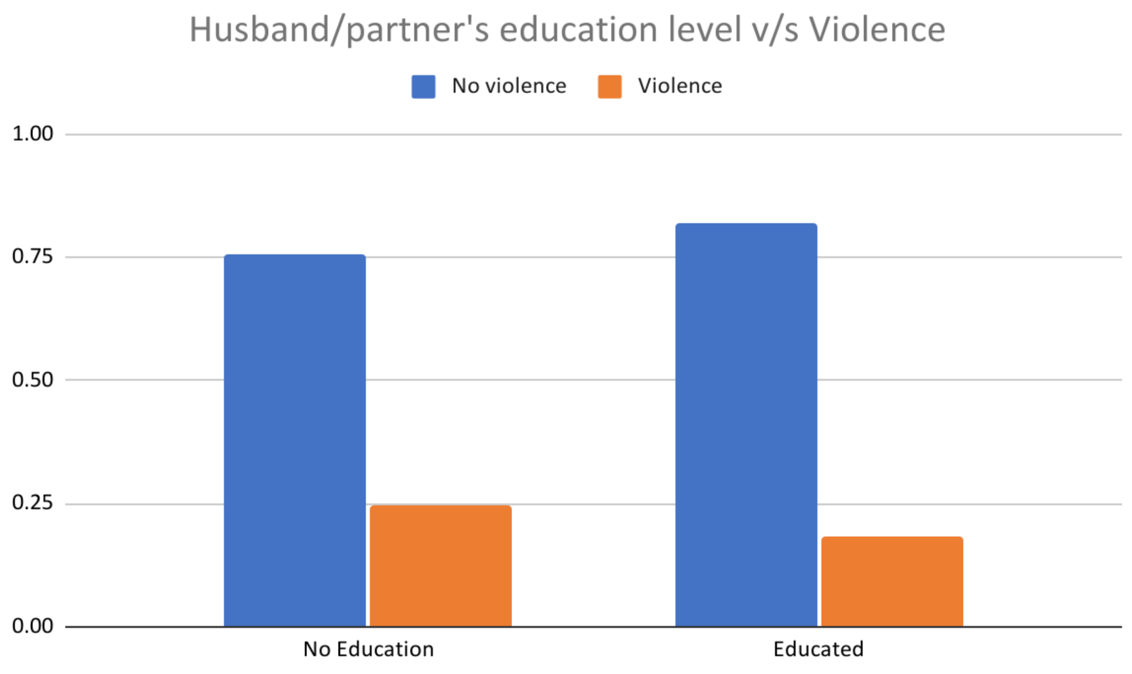
```
2-sample test for equality of proportions with continuity correction
data:  c(2276, 8267) out of c(9275, 45524)
X-squared = 201.41, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.05428299 0.07330564
sample estimates:
   prop 1    prop 2
0.2453908 0.1815965
```

- **Result**

As the p-value is less than 0.05 the null hypothesis is rejected and we conclude that:

1. Proportion of violence among those whose husbands are educated ≠ Proportion of violence among those whose husbands are not educated.
2. Husband/partner's education level has a significant effect on violence.
3. The proportion of violence among those whose husbands aren't educated is 0.2453 908.
4. The proportion of violence among those whose husbands are educated is 0.1815965.

➢ **Graph**



3.1.3. Violence v/s Frequency of Husband/partner being drunk

➢ **Hypothesis setup:**

Ho: Proportion of violence among those whose partner never drinks = Proportion of violence among those whose partner drinks.
H1: Proportion of violence among those whose partner never drinks ≠ Proportion of violence among those whose partner never drinks.

➢ **Proportion Test**

Frequency of husband/partner being drunk * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Frequency of husband/partner being drunk	Never Drinks	1011	249	1403
	Drinks	8375	3552	11927
Total		9386	3801	13187

P1 = x1/n1
P2 = x2/n2

Where,
P1 represents the proportion of violence among those whose partner never drinks.
x1 represents the number of victims who faced violence but their partner does not drink.
n1 represents the total number of those who never drink.
P2 represents the proportion of violence among those whose partner drinks.
x2 represents the number of victims who faced violence and their partner drinks.
n2 represents the total number of those who drink.

We will use R for calculating the proportions with the command being:

```
> prop.test(c(249,3552),c(1403,11927))
```

- **Output**

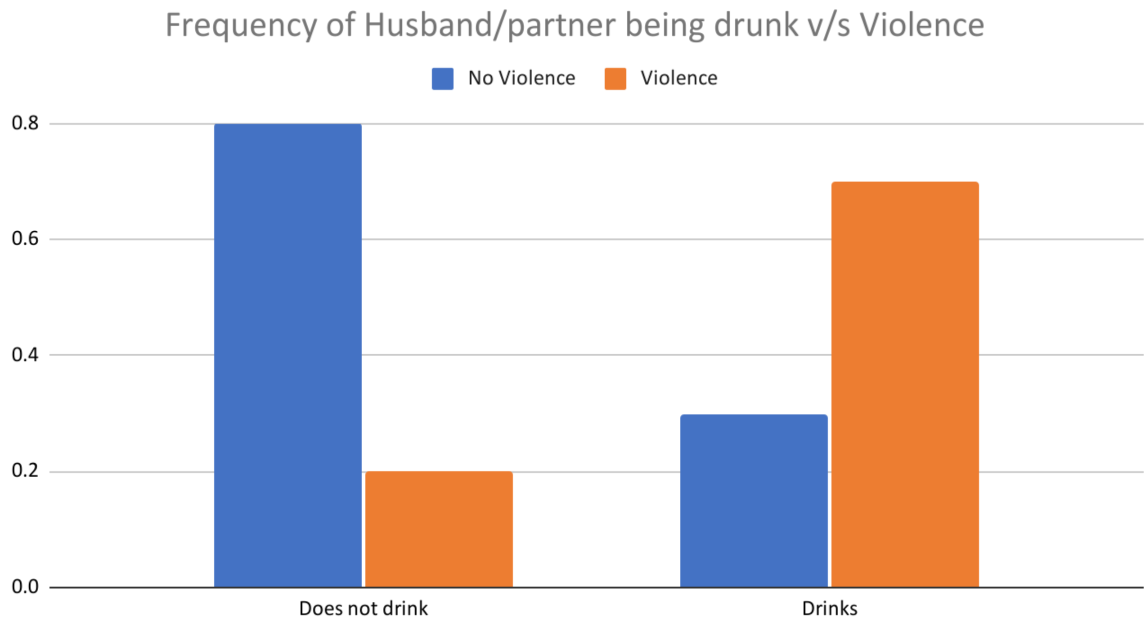
```
2-sample test for equality of proportions with continuity correction

data:  c(249, 3552) out of c(1403, 11927)
X-squared = 88.588, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.14234445 -0.09832525
sample estimates:
   prop 1    prop 2
0.1774768 0.2978117
```

- **Result**

As the p-value is less than 0.05, the null hypothesis is reject and we conclude that:

1. Proportion of violence among those whose partner never drinks ≠ Proportion of violence among those whose partner never drinks.
2. Drinking has a significant effect on violence.
3. The proportion of those who faced violence but their partner didn't drink is 0.1774768.
4. The proportion of those who faced violence and their partner drinks is 0.2978117.

➢ **Graph**



3.1.4. Violence v/s Respondent currently working

➢ **Hypothesis setup**

Ho: Proportion of violence among those who currently are not working = proportion of violence among those who currently are working.
H1: Proportion of violence among those who currently are not working ≠ proportion of violence among those who currently are working.

➢ **Proportion Test**

Respondent currently working * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Respondent currently working	No	32490	7085	44868
	Yes	11923	3495	18983
Total		44413	10580	63851

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who are not working currently.
x1 represents the number of victims who faced violence and are currently not working.
n1 represents the total number of people who are not working currently.
P2 represents the proportion of violence among those who are working currently.
x1 represents the number of victims who faced violence and are currently working.
n2 represents the total number of people who are working currently.

We will use R for calculating the proportion, with the command being:

```
> prop.test(c(7085,3495),c(44868,18983))
```

- **Output**

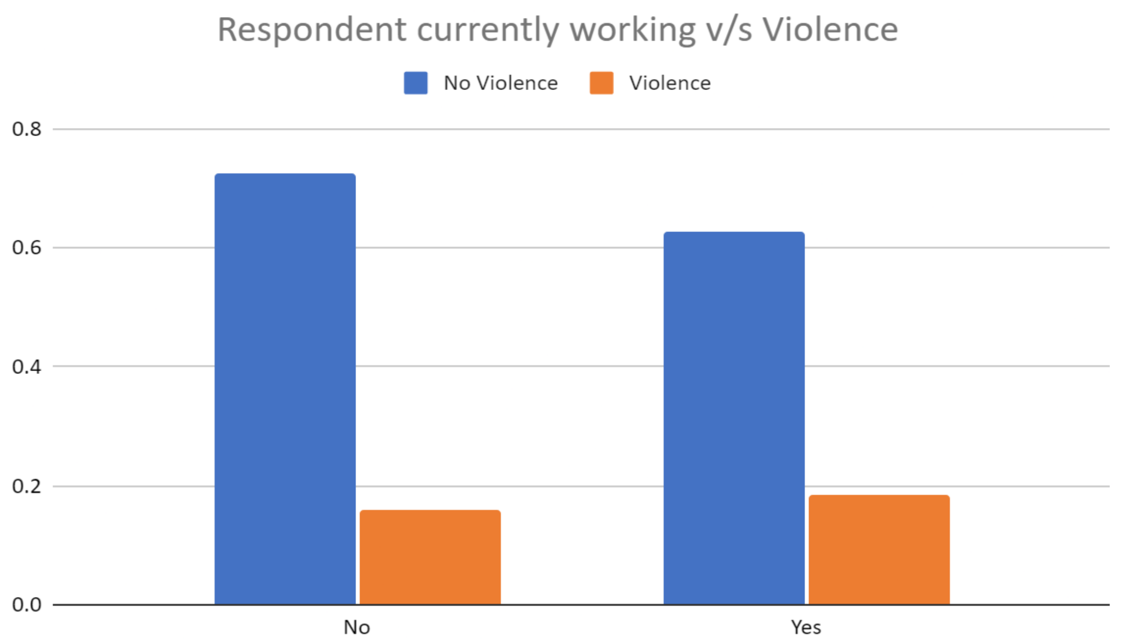
```
2-sample test for equality of proportions with continuity correction

data:  c(7085, 3495) out of c(44868, 18983)
X-squared = 66.069, df = 1, p-value = 4.353e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03270589 -0.01970303
sample estimates:
   prop 1    prop 2
0.1579076 0.1841121
```

- **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who currently are not working ≠ Proportion of violence among those who currently are working.
2. Working status has a significant effect on violence.
3. Proportion of those who faced violence and are not working currently is 0.1579076.
4. Proportion of those who faced violence and are currently working is 0.1841121.

➢ **Graph**



3.1.5. Violence v/s Respondent's Wealth Index

➢ **Hypothesis setup**

Ho: Proportion of violence among those who belong to below the poverty line = proportion of violence among those who belong to above the poverty line.
H1: Proportion of violence among those who belong to below the poverty line ≠ proportion of violence among those who belong to above the poverty line.

➢ **Proportion Test**

Wealth index combined * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Wealth index combined	Below Poverty Line	18281	5446	23727
	Above poverty line	26132	5134	31266
Total		44413	10580	54993

P1 = x1/n1
P2 = x2/n2

Where,
P1 represents the proportion of violence among those who belong to below the poverty line.
x1 represents the number of victims who faced violence and belonged to below the poverty line.
n1 represents the total cases of belonging to below the poverty line.
P2 represents the proportion of violence among those who belong to above the poverty line.
x2 represents the number of victims who faced violence and belonged to above the poverty line.
n2 represents the total cases of belonging to above the poverty line.

We will use R for calculating the proportion with the command being:

```
> prop.test(c(5446,5134),c(23727,31266))
```

- **Output**

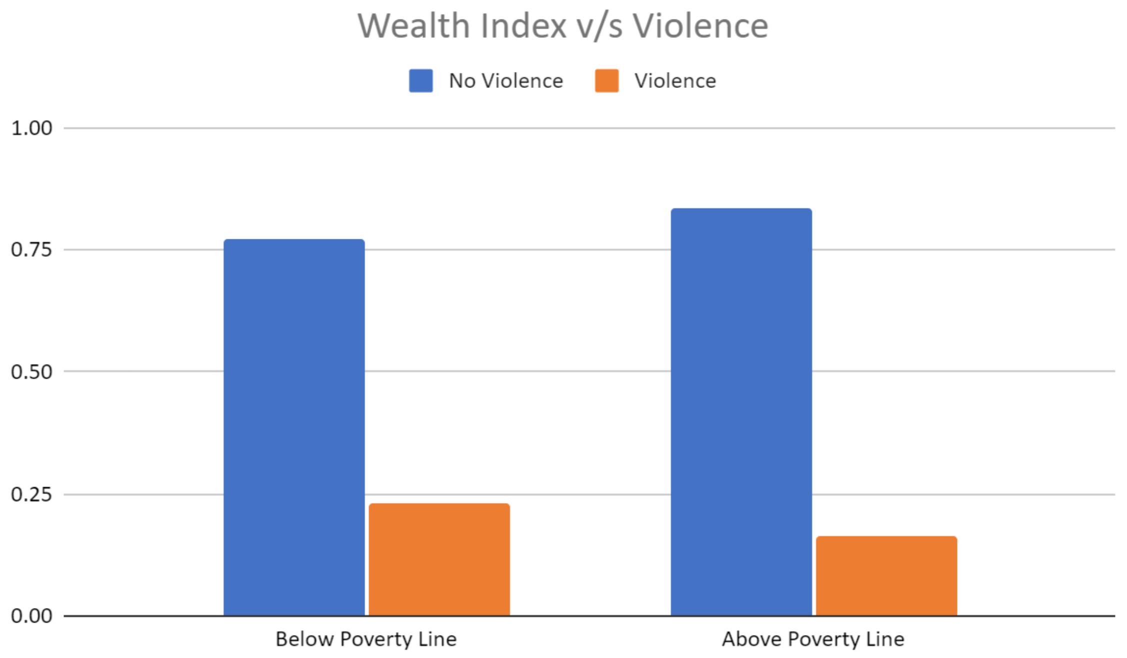
```
2-sample test for equality of proportions with continuity correction

data:  c(5446, 5134) out of c(23727, 31266)
X-squared = 370.06, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.05854166 0.07210557
sample estimates:
   prop 1    prop 2
0.2295275 0.1642039
```

- **Result**

As the p-value is less than 0.05 the null hypothesis is rejected and we conclude that:

1. Proportion of violence among those who belong to below the poverty line ≠ Proportion of violence among those who belong to above the poverty line.
2. The Wealth Index has a significant effect on violence.
3. The proportion of violence among those who belong to below the poverty line is 0.2295275.
4. The proportion of violence among those who belong to above the poverty line is 0.1642039.

➢ **Graph**



3.1.6. Violence v/s Type of Place of Residence

➢ **Hypothesis setup**

Ho: Proportion of violence among those who live in urban areas = proportion of violence among those who live in rural areas.
H1: Proportion of violence among those who live in urban areas ≠ proportion of violence among those who live in rural areas.

➢ **Proportion Test**

Type of place of residence * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Type of place of residence	Urban	11366	2289	13655
	Rural	33047	8291	41338
Total		44413	10580	54993

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who live in urban areas.
x1 represents the number of victims who faced violence and live in urban areas.
n1 represents the total number of people who live in urban areas.
P2 represents the proportion of violence among those who live in rural areas.
x1 represents the number of victims who faced violence and live in rural areas.
n2 represents the total number of people who live in rural areas.

We will use R for calculating the proportion, with the command being:
> prop.test(c(2289,8291),c(13655,41338))

- **Output**

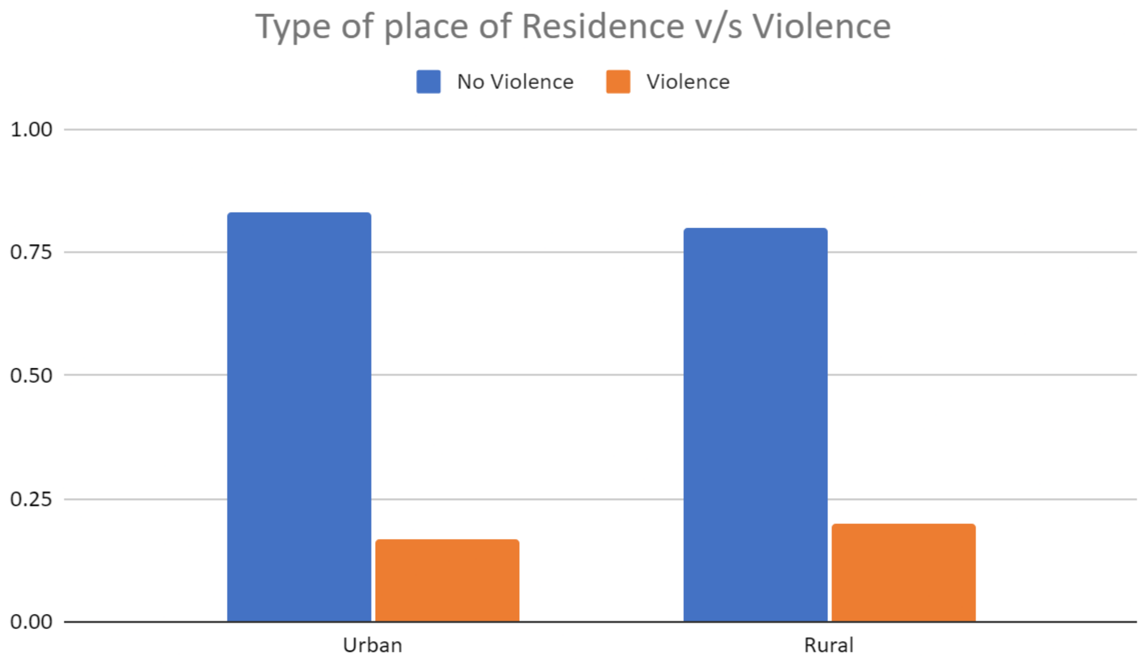
2-sample test for equality of proportions with continuity correction
data: c(2289, 8291) out of c(13655, 41338)
X-squared = 71.447, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
-0.04034276 -0.02552757
sample estimates:
prop 1 prop 2
0.1676309 0.2005661

- **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who live in urban areas ≠ Proportion of violence among those who live in rural areas.
2. Type of place of residence has a significant effect on violence.
3. Proportion of those who faced violence and lived in urban areas is 0.1676309.
4. Proportion of those who faced violence and lived in rural areas is 0.2005661.

➢ **Graph**



3.1.7. Violence v/s Having children or not

➢ **Hypothesis setup**

Ho: Proportion of violence among those who have children = proportion of violence among those who don't have children.
H1: Proportion of violence among those who have children ≠ proportion of violence among those who don't have children.

➢ **Proportion Test**

Having children or not * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Having children or not	No children	3909	618	4527
	Have children	40504	9962	50466
Total		44413	10580	54993

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who don't have children.
x1 represents the number of victims who faced violence but didn't have any children.
n1 represents the total number of people who don't have children.
P2 represents the proportion of violence among those who have children.
x1 represents the number of victims who faced violence and had children.
n2 represents the total number of people who have children

We will use R for calculating the proportion, with the command being:
> prop.test(c(618,9962),c(4527,50466))

- **Output**

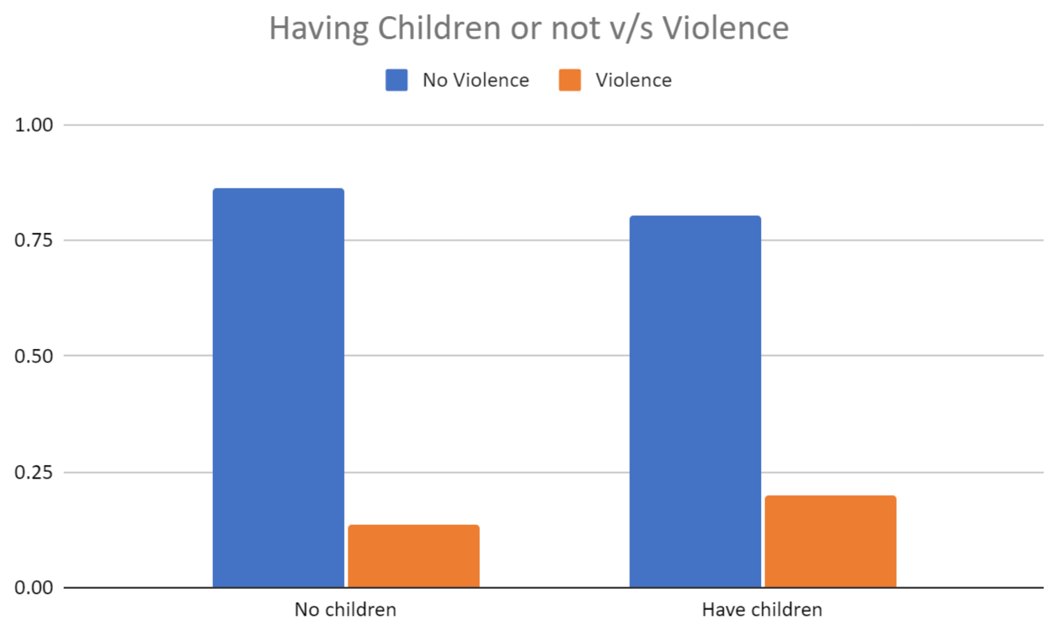
2-sample test for equality of proportions with continuity correction
data: c(618, 9962) out of c(4527, 50466)
X-squared = 98.727, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
-0.07159347 -0.05017850
sample estimates:
prop 1 prop 2
0.1365142 0.1974002

- **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who have children ≠ Proportion of violence among those who don't have children.
2. Having children has a significant effect on violence.
3. Proportion of those who faced violence and didn't have any children is 0.1365142.
4. Proportion of those who faced violence and did have children is 0.1974002.

➢ **Graph**



3.1.8. Violence v/s Pregnancy Loss

➢ **Hypothesis setup**

Ho: Proportion of violence among those who had pregnancy loss = proportion of violence among those who had abortion.
H1: Proportion of violence among those who had pregnancy loss ≠ proportion of violence among those who had abortion.

➢ **Proportion Test**

Pregnancy end in miscarriage, abortion, or stillbirth * Violence Crosstabulation				
Count				
		Violence		
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Total
Pregnancy end in miscarriage, abortion, or stillbirth	Pregnancy Loss	1786	467	2253
	Abortion	603	195	798
Total		2389	662	3051

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who had pregnancy loss.
x1 represents the number of victims who faced violence and had pregnancy loss.
n1 represents the total number of people who had pregnancy loss.
P2 represents the proportion of violence among those who had abortion.
x1 represents the number of victims who faced violence and had abortion.
n2 represents the total number of people who had abortion.

We will use R for calculating the proportion, with the command being:

```
>  prop.test(c(467,195),c(2253,798))
```

- **Output**

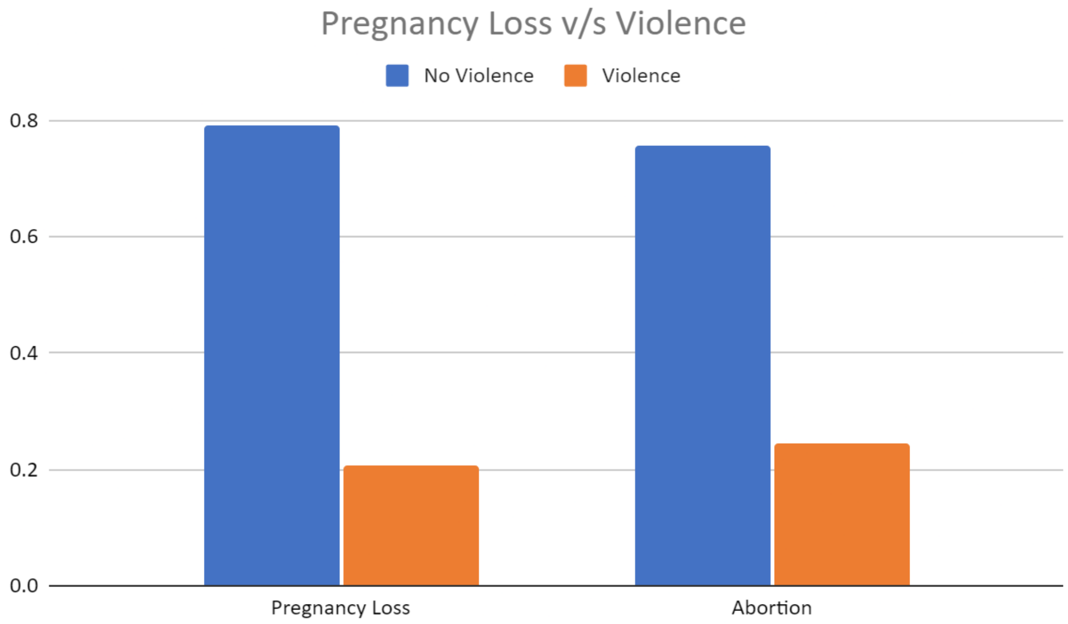
```
2-sample test for equality of proportions with continuity correction

data:  c(467, 195) out of c(2253, 798)
X-squared = 4.5535, df = 1, p-value = 0.03285
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.072121373 -0.002042065
sample estimates:
   prop 1    prop 2
0.2072792 0.2443609
```

- **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who had pregnancy loss ≠ Proportion of violence among those who had abortion.
2. Pregnancy loss has a significant effect on violence.
3. Proportion of those who faced violence and had pregnancy loss is 0.2072792.
4. Proportion of those who faced violence and had abortion is 0.2443609.

➢ **Graph**



3.1.9. Violence v/s Related to current husband prior to marriage

➢ **Hypothesis setup**

Ho: Proportion of violence among those who were related to their husband prior to marriage = Proportion of violence among those who weren't related to their husband prior to marriage.
H1: Proportion of violence among those who were related to their husband prior to marriage ≠ Proportion of violence among those who weren't related to their husband prior to marriage.

➢ **Proportion Test**

Related to current husband prior to marriage in any way * Violence Crosstabulation				
Count				
		Violence		Total
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	
Related to current husband prior to marriage in any way	No	39816	9087	56401
	Yes	4597	1493	7450
Total		44413	10580	63851

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who weren't related to their current husband prior to marriage.
x1 represents the number of victims who faced violence but weren't related to their current husband prior to marriage.
n1 represents the total number of people who weren't related to their current husband prior to marriage.
P2 represents the proportion of violence among those who were related to their current husband prior to marriage.
x1 represents the number of victims who faced violence and were related to their current husband prior to marriage.
n2 represents the total number of people who were related to their current husband prior to marriage.

We will use R for calculating the proportion, with the command being:
> prop.test(c(9087,1493),c(56401,7450))

- ➢ **Output**

2-sample test for equality of proportions with continuity correction

data: c(9087, 1493) out of c(56401, 7450)
X-squared = 73.195, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
-0.04894735 -0.02962969
sample estimates:
prop 1 prop 2
0.1611142 0.2004027

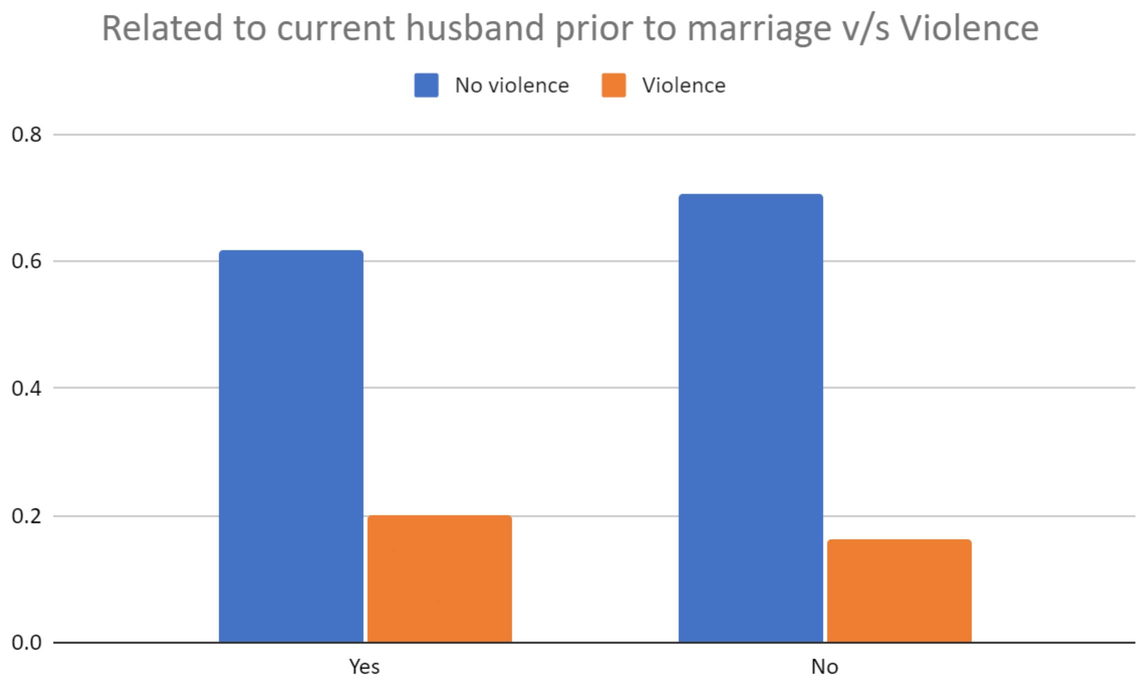
- ➢ **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who were related to their current husband prior to marriage ≠ Proportion of violence among those who weren't related to their current husband prior to marriage.
2. Being related to the current husband prior to marriage has a significant effect on violence.

3. Proportion of those who faced violence and weren't related to their current husband prior to marriage is 0.1611142
4. Proportion of those who faced violence but were related to their current husband prior to marriage is 0.2004027.

➢ **Graph**



3.1.10. Violence v/s Can you say no to your husband if you do not want to have sexual intercourse

➢ **Hypothesis setup**

Ho: Proportion of violence among those who said no to intercourse to their husband = Proportion of violence among those who didn't say no to intercourse to their husband.
H1: Proportion of violence among those who said no to intercourse to their husband ≠ Proportion of violence among those who didn't say no to intercourse to their husband.

- **Proportion Test**

Can you say no to your husband if you do not want to have sexual intercourse? * Violence Crosstabulation				
Count				
		Violence		Total
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	
Can you say no to your husband if you do not want to have sexual intercourse?	No	7793	1840	11511
	Yes	34613	8247	48969
Total		42406	10087	60480

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who didn't say no to intercourse to their husband.
x1 represents the number of victims who faced violence and didn't say no to intercourse to their husband.
n1 represents the total number of people who didn't say no to intercourse to their husband.
P2 represents the proportion of violence among those who said no to intercourse to their husband.
x1 represents the number of victims who faced violence and said no to intercourse to their husband.
n2 represents the total number of people who said no to intercourse to their husband.

We will use R for calculating the proportion, with the command being:

```
> prop.test(c(1840,8247),c(11511,48969))
```

➢ **Output**

2-sample test for equality of proportions with continuity correction

data: c(1840, 8247) out of c(11511, 48969)
X-squared = 4.8592, df = 1, p-value = 0.0275
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.016089409 -0.001041724
sample estimates:
   prop 1   prop 2
0.1598471 0.1684127

➢ **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who said no to intercourse to their husband ≠ Proportion of violence among those who didn't say no to intercourse to their husband.
2. Proportion of those who faced violence and didn't say no to intercourse to their husband is 0.1598471.
3. Proportion of those who faced violence and did say no to intercourse to their husband is 0.1684127

➢ **Graph**

3.1.11. Violence v/s Caste

➢ **Hypothesis setup**

Ho: proportion of violence among those who belong to lower caste = proportion of violence among those who belong to upper caste.
H1: proportion of violence among those who belong to lower caste ≠ proportion of violence among those who belong to upper caste.

➢ **Proportion Test**

Belong to a scheduled caste, a scheduled tribe, other backward class * Violence Crosstabulation

Count

		Violence		Total
		No violence occurred	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	
Belong to a scheduled caste, a scheduled tribe, other backward class	Lower Caste(SC/ST /OBC)	33145	8626	41771
	Upper Caste(General)	8745	1446	10191
Total		41890	10072	51962

P1 = x1/n1
P2 = x2/n2

Where,

P1 represents the proportion of violence among those who belong to the lower caste.
x1 represents the number of victims who faced violence and belonged to the lower caste.
n1 represents the total number of people who belong to lower caste.
P2 represents the proportion of violence among those who belong to the upper caste.
x1 represents the number of victims who faced violence and belonged to the upper caste.
n2 represents the total number of people who belong to the upper caste.

We will use R for calculating the proportion, with the command being:

```
>  prop.test(c(8626,1446),c(41771,10191))
```

- **Output**

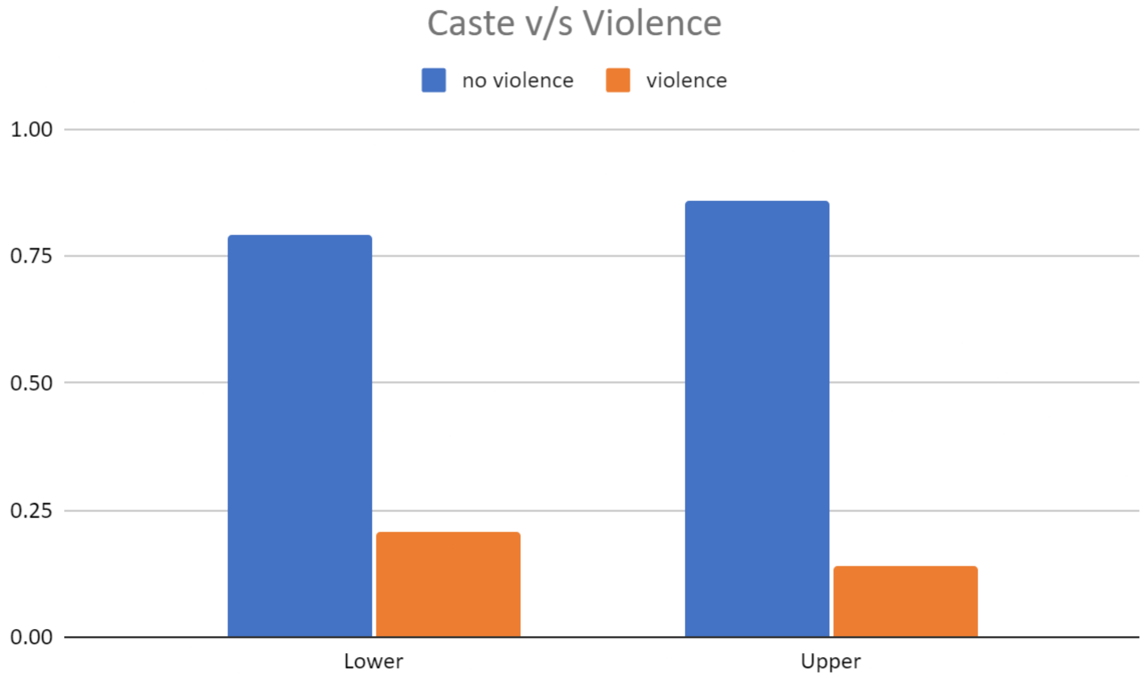
```
2-sample test for equality of proportions with continuity correction

data:  c(8626, 1446) out of c(41771, 10191)
X-squared = 218.49, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.05674794 0.07248607
sample estimates:
   prop 1    prop 2
0.2065069 0.1418899
```

- **Result**

As the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that:

1. Proportion of violence among those who belong to lower caste ≠ Proportion of violence among those who belong to the upper caste.
2. Caste has a significant effect on violence.
3. Proportion of those who faced violence and belonged to lower caste is 0.2065069
4. Proportion of those who faced violence and belonged to the upper caste is 0.1418899.

- **Graph**



3.2 Chi-square Test

The chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, which means that it does not require any assumptions about the distribution of the data being analyzed. This test is widely used in social and health sciences, as well as in market research and quality control. The chi-square test is based on the comparison of the observed frequencies of different categories with the expected frequencies, assuming that there is no association between the variables. If the observed frequencies differ significantly from the expected frequencies, then the chi-square test indicates that there is a significant association between the variables. A chi-square test is a valuable tool in research and data analysis, providing insights into the relationships between different categorical variables.

The chi-square test is set up by first defining the null hypothesis and the alternative hypothesis. The null hypothesis states that there is no association between the two categorical variables being compared, while the alternative hypothesis states that there is a significant association between the variables.

Next, a contingency table is created, which shows the observed frequencies of the different categories for each of the variables being analyzed. The contingency table is a two-way table, with the rows representing one variable and the columns representing the other variable. Once the contingency table has been created, the expected frequencies are calculated for each cell of the table, assuming that there is no association between the variables. The expected frequency for each cell is calculated by multiplying the marginal totals for the row and column that the cell belongs to and then dividing by the total sample size.

The chi-square statistic is then calculated by summing the squared differences between the observed and expected frequencies for each cell and dividing by the expected frequency for that cell. The resulting value is compared to a chi-square distribution table with degrees of freedom equal to the product of the number of rows minus one and the number of columns minus one. If the calculated chi-square value is greater than the critical value from the chi-square distribution table, then the null hypothesis is rejected, and it is concluded that there is a significant association between the two variables being compared.

3.2.1. Violence v/s Respondent's education level

- **Hypothesis set up:**

H_o: There is no dependency between the respondent's education level and violence.
H_1: There is a dependency between the respondent's education level and violence.

Highest educational level * Violence Crosstabulation

Count

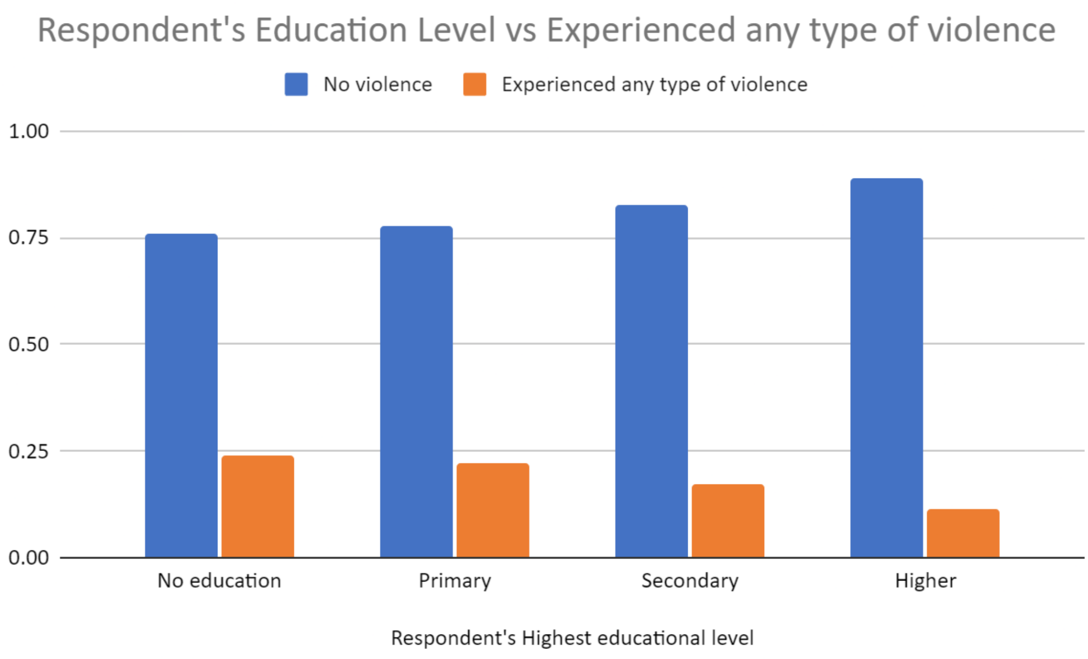
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Highest educational level	No education	11601	3687	2038	940	517	18783
	Primary	6134	1742	783	430	213	9302
	Secondary	21028	4436	2026	934	519	28943
	Higher	5650	715	269	119	70	6823
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1355.078[a]	12	<.001
Likelihood Ratio	1400.809	12	<.001
Linear-by-Linear Association	1077.065	1	<.001
N of Valid Cases	63851		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 140.95.

➢ **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that the respondent's education level has a significant effect on violence. As the respondent's education level increases, the respondent is facing less and less violence.

3.2.2. Violence v/s Husband/partner’s education level

- **Hypothesis set up:**

H_o: There is no dependency between the husband/partner’s education level and violence.
H_1: There is a dependency between the husband/partner’s education level and violence.

Husband/partner's education level * Violence Crosstabulation

Count

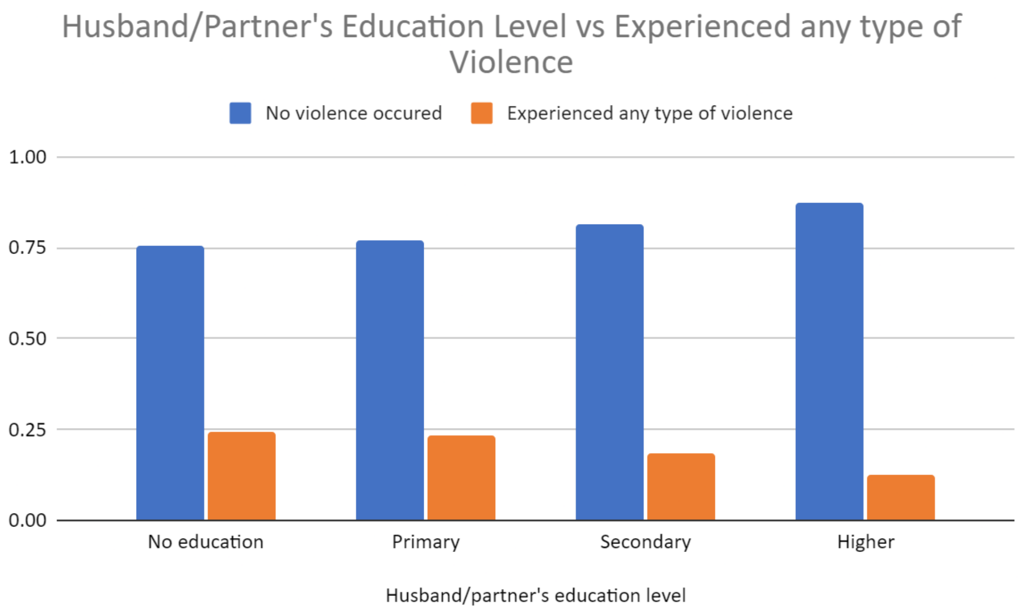
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Husband/partner's education level	No education	6999	2276	1278	672	403	11628
	Primary	5911	1782	886	445	244	9268
	Secondary	24230	5456	2513	1127	576	33902
	Higher	7116	1029	417	169	88	8819
	Don't know	157	37	22	10	8	234
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1337.583[a]	16	<.001
Likelihood Ratio	1346.537	16	<.001
Linear-by-Linear Association	988.172	1	<.001
N of Valid Cases	63851		

a. 1 cells (4.0%) have expected count less than 5. The minimum expected count is 4.83.

➢ Graph



➢ Interpretation:

From the graph, we can clearly see that the husband's education level has a significant effect on violence. As the husband's education level increases, the respondent is facing less and less violence.

3.2.3. Violence v/s Frequency of Husband/partner being drunk

➢ Hypothesis set up:

H_0: There is no dependency between the frequency of Husband/Partner being drunk and violence.
H_1: There is a dependency between the frequency of Husband/Partner being drunk and violence.

Frequency of husband/partner being drunk * Violence Crosstabulation

Count

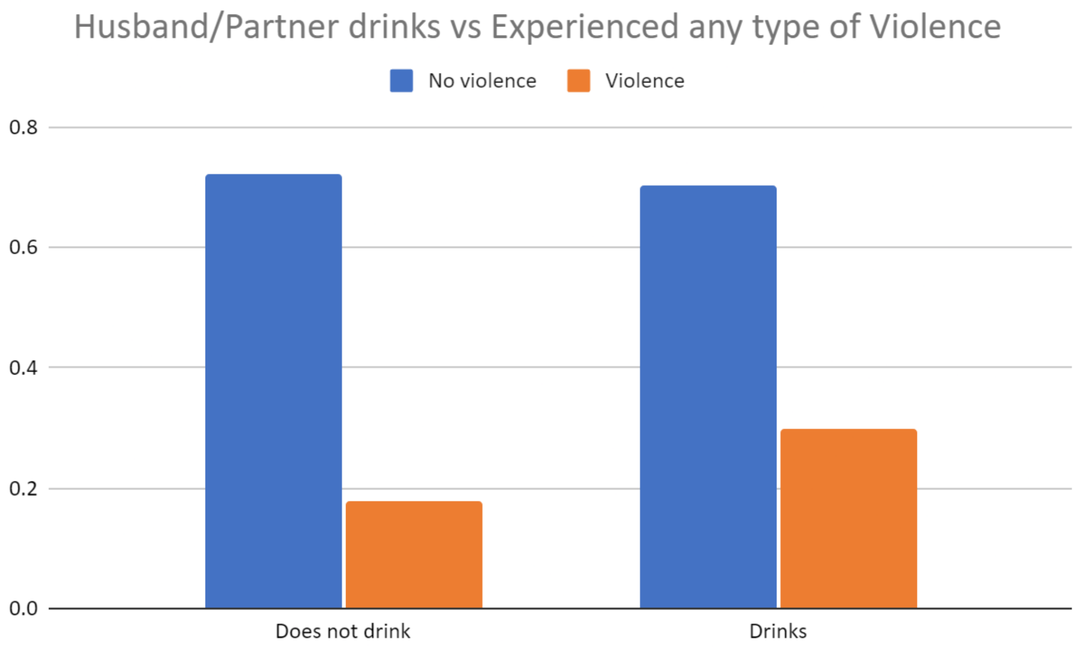
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Frequency of husband/partner being drunk	Never	1011	249	93	37	13	1403
	Often	1145	596	586	517	468	3312
	Sometimes	7230	2956	1687	824	420	13117
Total		9386	3801	2366	1378	901	17832

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1478.230[a]	8	.000
Likelihood Ratio	1319.606	8	<.001
Linear-by-Linear Association	67.015	1	<.001
N of Valid Cases	17832		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 70.89.

➢ Graph



➢ Interpretation:

From the graph, we can clearly see that the husband's drinking status has a significant effect on violence. If the husband drinks, the respondent faces more violence.

3.2.4. Violence v/s Respondent's Working Status

➢ Hypothesis set up:

H_0: There is no dependency between the respondent's working status and violence.
H_1: There is a dependency between the respondent's working status and violence.

Respondent currently working * Violence Crosstabulation

Count

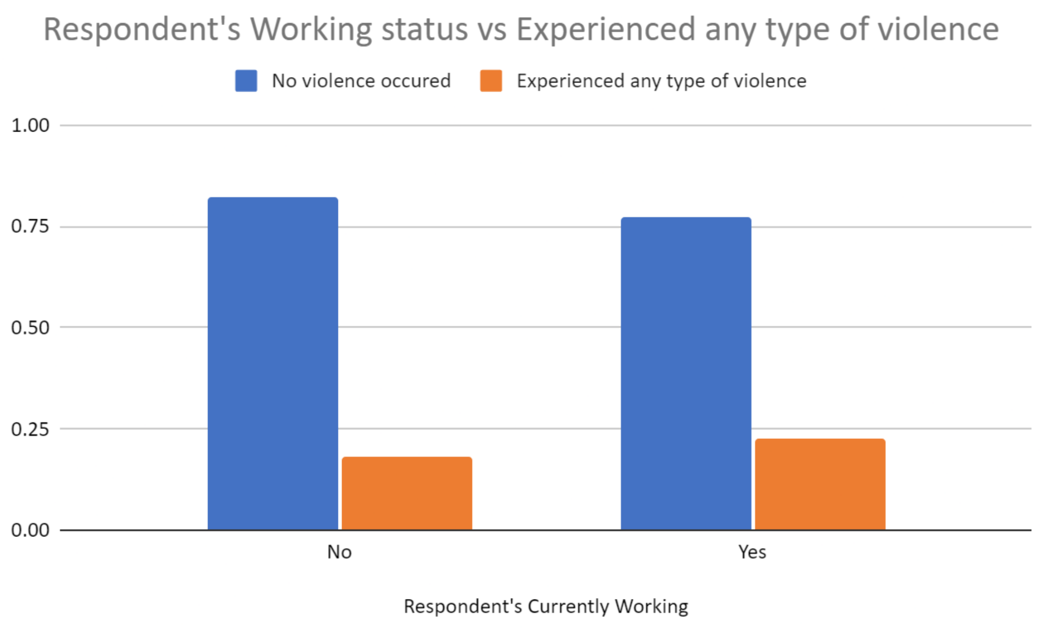
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Respondent currently working	No	32490	7085	3179	1391	723	44868
	Yes	11923	3495	1937	1032	596	18983
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	736.748[a]	4	<.001
Likelihood Ratio	708.744	4	<.001
Linear-by-Linear Association	727.881	1	<.001
N of Valid Cases	63851		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 392.14.

➢ **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that the respondent's working status has a significant effect on violence. If the respondent works, she faces more violence.

3.2.5. Violence v/s Respondent's Wealth Index

➢ **Hypothesis set up:**

H_0: There is no dependency between the respondent's wealth index and violence.

H_1: There is a dependency between the respondent's wealth index and violence.

Wealth index combined * Violence Crosstabulation

Count

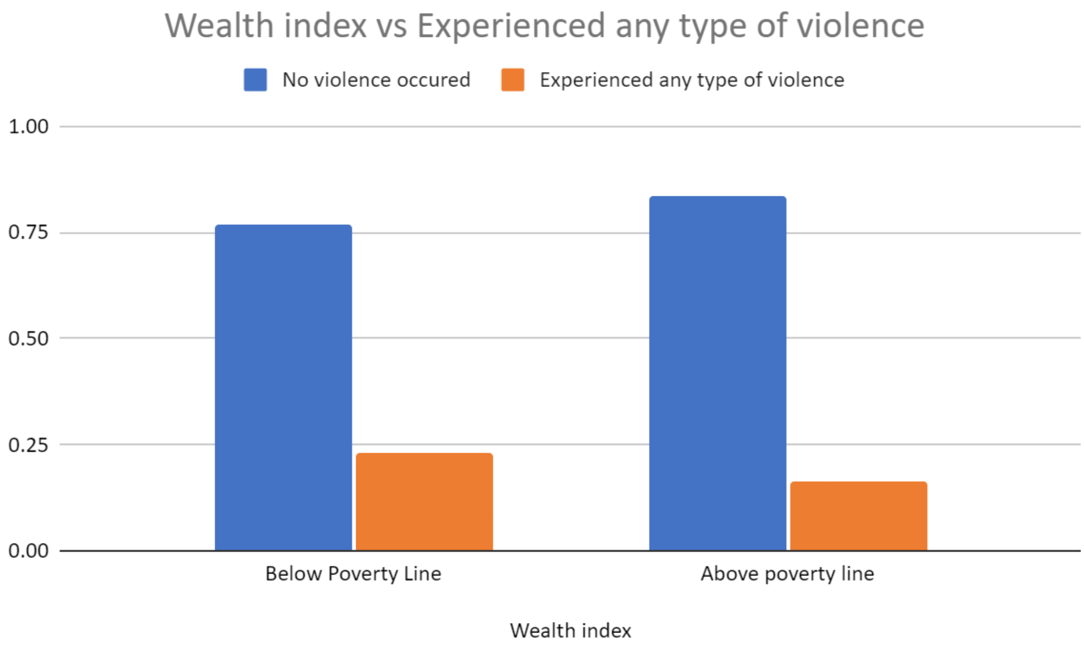
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Wealth index combined	Poorest	8723	2845	1452	704	423	14147
	Poorer	9558	2601	1310	693	335	14497
	Middle	9176	2248	1058	497	284	13263
	Richer	8737	1784	818	336	178	11853
	Richest	8219	1102	478	193	99	10091
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1337.472[a]	16	<.001
Likelihood Ratio	1393.010	16	<.001
Linear-by-Linear Association	1108.830	1	<.001
N of Valid Cases	63851		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 208.45.

➢ **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that the wealth index has a significant effect on violence. If the respondent lives below the poverty line, she experiences more violence.

3.2.6. Violence v/s Type of place of residence

➢ **Hypothesis set up:**

H_o: There is no dependency between the type of place of residence and Violence.
H_1: There is a dependency between the type of place of residence and Violence.

Type of place of residence * Violence Crosstabulation

Count

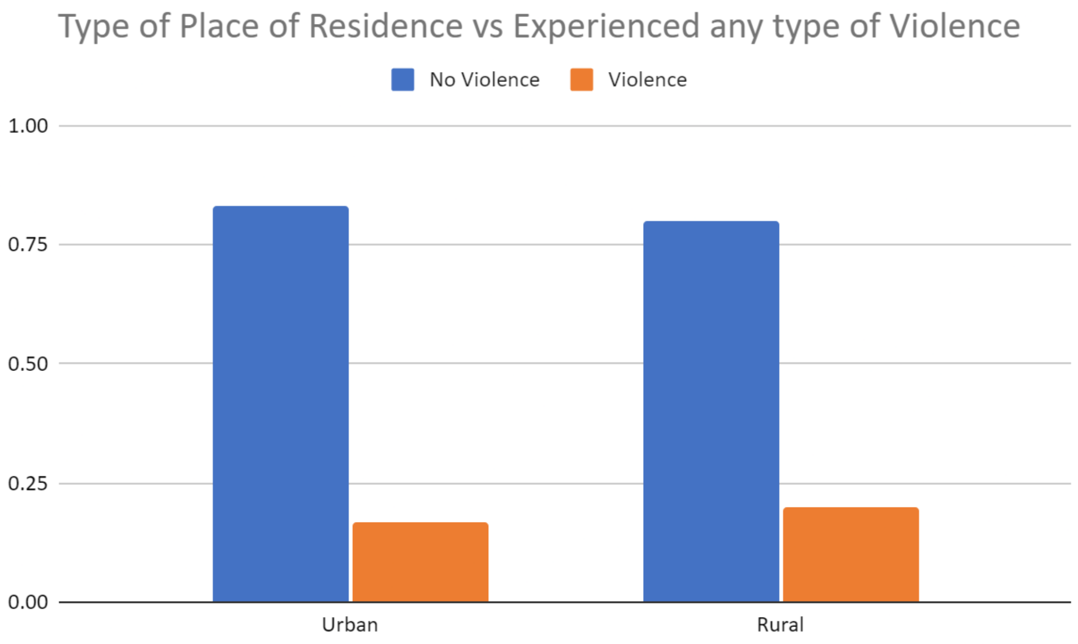
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Type of place of residence	Urban	11366	2289	1082	456	295	15488
	Rural	33047	8291	4034	1967	1024	48363
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	151.092[a]	4	<.001
Likelihood Ratio	154.769	4	<.001
Linear-by-Linear Association	114.443	1	<.001
N of Valid Cases	63851		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 319.94.

- **Graph**



- **Interpretation:**

From the graph, we can clearly see that the type of place of residence has a significant effect on violence. If the respondent lives in a rural area, she faces more violence.

3.2.7. Violence v/s Number of living children

- **Hypothesis set up:**

H_o: There is no dependency between the number of living children and violence.
H_1: There is a dependency between the number of living children and violence.

Number of living children * Violence Crosstabulation

Count

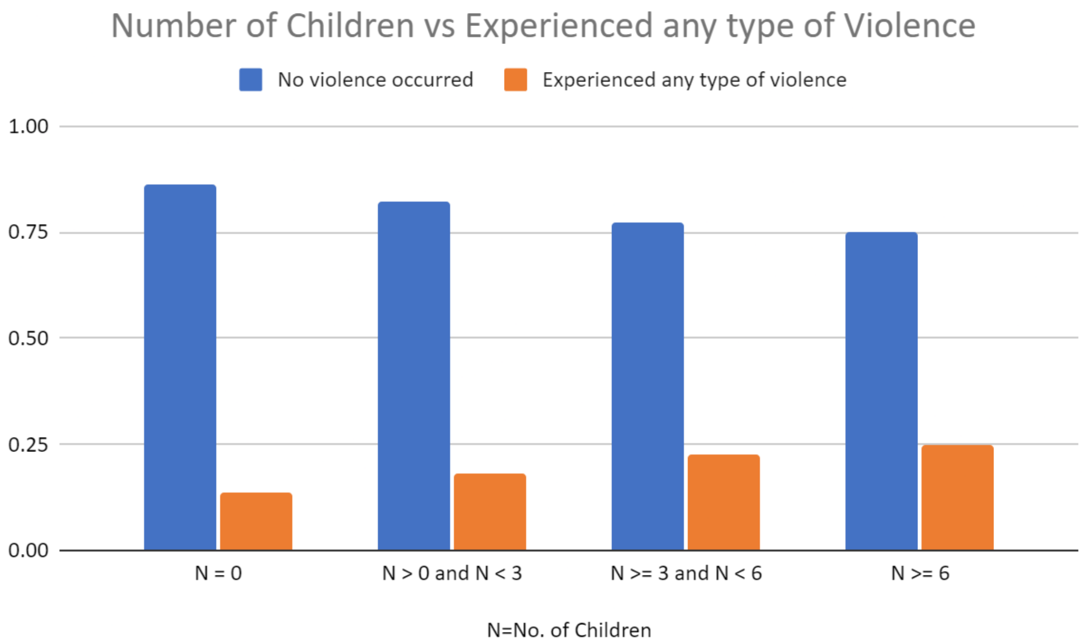
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Number of living children	0	3909	618	315	175	98	5115
	1	9033	1783	809	410	233	12268
	2	16875	3876	1854	844	451	23900
	3	8659	2441	1232	542	290	13164
	4	3672	1142	542	285	127	5768
	5	1413	439	215	101	68	2236
	6	531	171	90	40	30	862
	7	206	73	44	16	16	355
	8	74	31	9	8	4	126
	9	24	5	6	2	2	39
	10	12	1	0	0	0	13
	11	4	0	0	0	0	4
	12	1	0	0	0	0	1
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	554.068[a]	48	<.001
Likelihood Ratio	551.917	48	<.001
Linear-by-Linear Association	327.793	1	<.001
N of Valid Cases	63851		

a. 19 cells (29.2%) have expected count less than 5. The minimum expected count is .02.

- ➢ **Graph**



- ➢ **Interpretation:**

From the graph, we can clearly see that the number of children has a significant effect on violence. As the number of children increases, the respondent faces more violence.

3.2.8. Violence v/s Belong to a scheduled caste, a scheduled tribe, or backward class

- **Hypothesis set up:**

H_0: There is no dependency between a respondent belonging to a scheduled caste, scheduled tribe, or backward class and violence.
H_1: There is a dependency between the respondent's belonging to a scheduled caste, scheduled tribe, or backward class and violence.

Belong to a scheduled caste, a scheduled tribe, other backward class * Violence Crosstabulation

Count

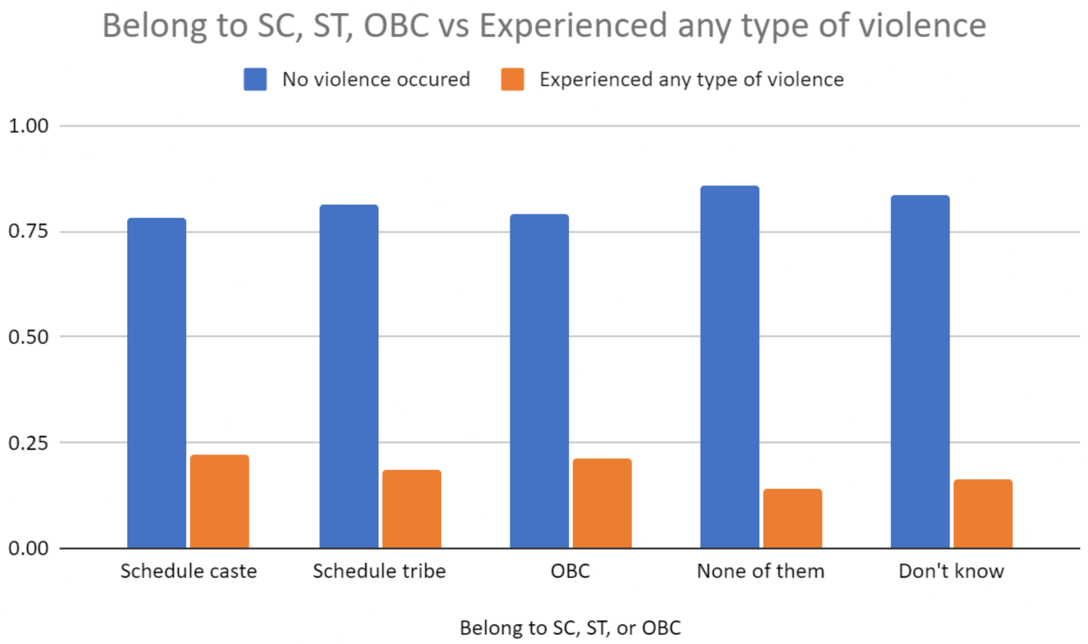
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Belong to a scheduled caste, a scheduled tribe, other backward class	Schedule caste	7857	2208	1169	577	353	12164
	Schedule tribe	8657	1970	939	420	234	12220
	OBC	16631	4448	2089	993	500	24661
	None of them	8745	1446	680	301	178	11350
	Don't know	258	51	30	20	8	367
Total		42148	10123	4907	2311	1273	60762

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	543.334[a]	16	<.001
Likelihood Ratio	553.072	16	<.001
Linear-by-Linear Association	211.907	1	<.001
N of Valid Cases	60762		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.69.

➢ **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that caste has a significant effect on violence. The chance of experiencing violence will be higher in Schedule Caste as compared to Schedule Tribe or OBC.

3.2.9. Violence v/s Pregnancy ending in miscarriage, abortion, or stillbirth

➢ **Hypothesis set up:**

H_0: There is no dependency between pregnancy ending in miscarriage, abortion, or stillbirth and violence.
H_1: There is a dependency between pregnancy ending in miscarriage, abortion, or stillbirth and violence.

Pregnancy end in miscarriage, abortion, or stillbirth * Violence Crosstabulation

Count

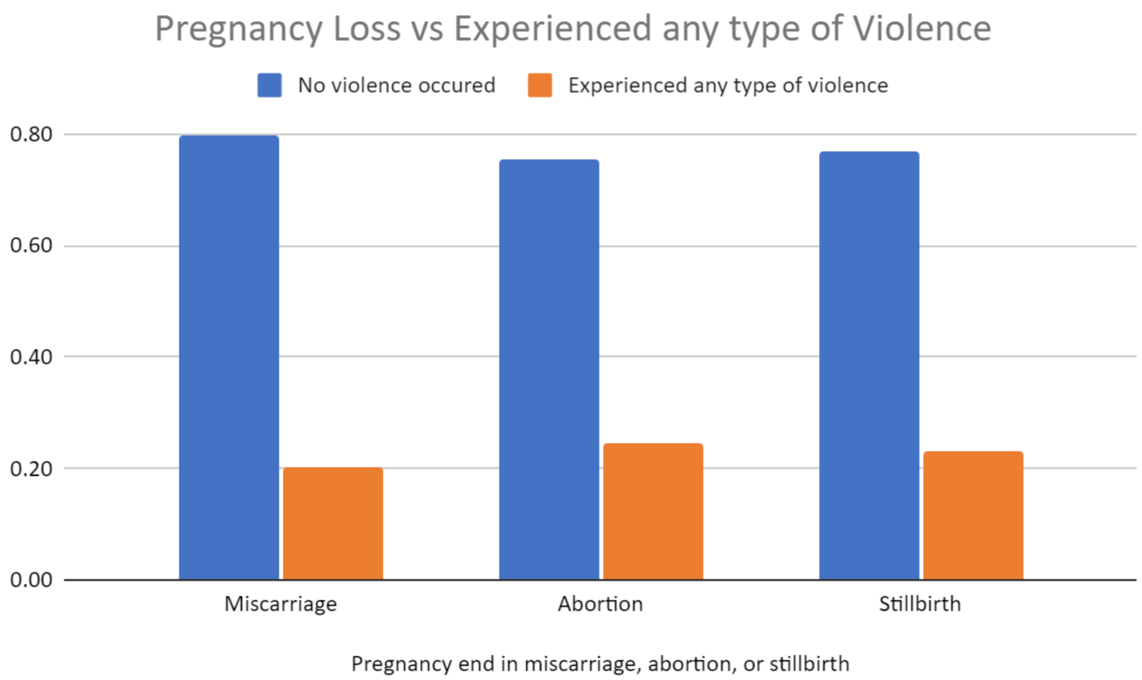
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Pregnancy end in miscarriage, abortion, or stillbirth	Miscarriage	1585	406	186	110	53	2340
	Abortion	603	195	113	51	26	988
	Stillbirth	201	61	28	10	4	304
Total		2389	662	327	171	83	3632

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	20.448[a]	8	.009
Likelihood Ratio	20.431	8	.009
Linear-by-Linear Association	2.013	1	.156
N of Valid Cases	3632		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.95.

- **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that pregnancy loss has a significant effect on violence. Among miscarriage, abortion, and stillbirth, if the pregnancy loss is due to abortion, the respondent faces most violence.

3.2.10. Violence v/s Respondent's Relationship to Current Husband Prior to Marriage in Any Way

➢ **Hypothesis set up:**

H_0: There is no dependency between the respondent's relationship with the current husband prior to marriage in any way and violence.
H_1: There is a dependency between the respondent's relationship with the current husband prior to marriage in any way and violence.

Related to current husband prior to marriage in any way * Violence Crosstabulation

Count

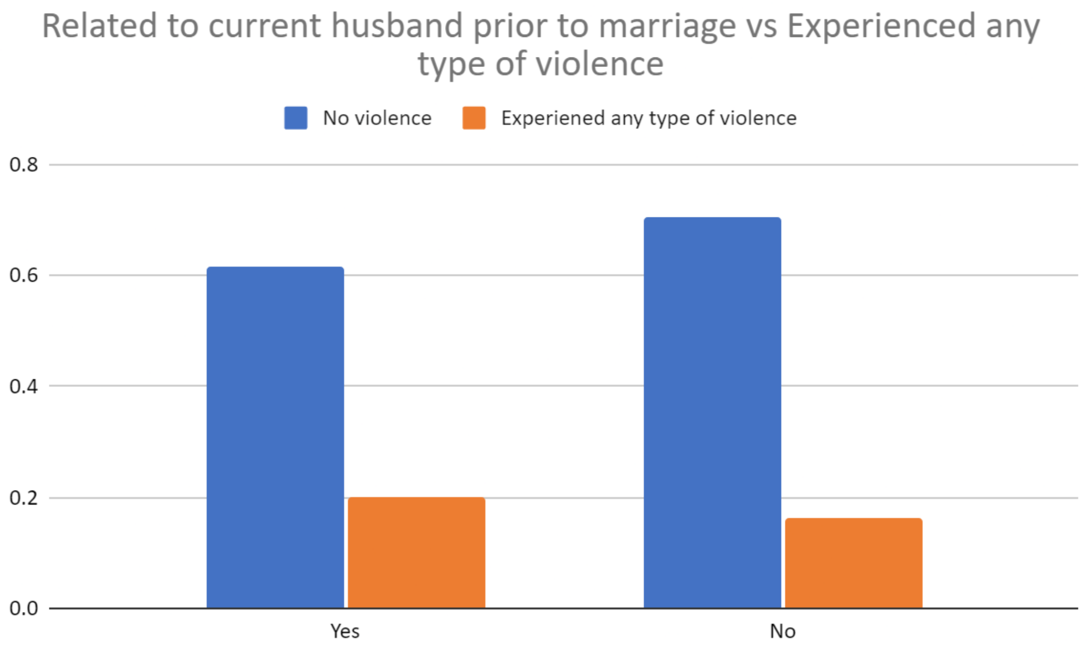
		Violence					
		No violence occured	Experienced any type of violence whether Emotional, Sexual, Less Severe, or Severe	Experienced any two types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced any three types of violence: Emotional, Sexual, Less Severe, or Severe	Experienced all types of violence including Emotional, Sexual, Less Severe and Severe	Total
Related to current husband prior to marriage in any way	No	39816	9087	4351	2026	1121	56401
	Yes	4597	1493	765	397	198	7450
Total		44413	10580	5116	2423	1319	63851

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	256.262[a]	4	<.001
Likelihood Ratio	245.532	4	<.001
Linear-by-Linear Association	212.994	1	<.001
N of Valid Cases	63851		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 153.90.

➢ **Graph**



➢ **Interpretation:**

From the graph, we can clearly see that relationships with husbands prior to marriage have a significant effect on violence. If the respondent was related to her current husband prior to marriage, she faces more violence.

3.2.11. Results

Criterion	Critical Value (5% level of significance)	Chi-square Statistic	P-value	Decision
Independence of Violence and Respondent's education level	5.226	1355.078	<.001	Reject null hypothesis
Independence of Violence and Husband's education level	7.962	1337.583	<.001	Reject null hypothesis

Independence of Violence and Frequency of Husband/partner being drunk	2.733	1478.230	0.001	Reject null hypothesis
Independence of Violence and Respondent working	0.711	736.748	<.001	Reject null hypothesis
Independence of Violence and Wealth index	7.962	1337.472	<.001	Reject null hypothesis
Independence of Violence and Type of Place of residence	0.711	151.092	<0.001	Reject null hypothesis
Independence of Violence and Number of living children	33.098	554.068	<.001	Reject null hypothesis
Independence of Violence and Respondent belonging to SC, ST, or OBC	7.962	543.334	<.001	Reject null hypothesis
Independence of Violence and Pregnancy end in miscarriage, abortion, or stillbirth	2.733	20.448	0.009	Reject null hypothesis
Independence of Violence and Relation to current husband prior to Marriage	0.711	256.262	<.001	Reject null hypothesis

3.3. Empirical Analysis of the Odds Ratio

It is a statistical measure used to compare the odds of an event occurring in one group to the odds of the same event occurring in another group. The odds ratio is calculated by dividing the odds of an event occurring in one group by the odds of the same event occurring in another group. If the odds ratio is greater than 1, it indicates that the event is more likely to occur in the first group, while an odds ratio of less than 1 indicates that the event is more likely to occur in the second group. An odds ratio of 1 indicates no difference in the odds of the event occurring in either group.

One limitation of the odds ratio is that it can be difficult to interpret. The odds ratio can be complex, and the results may be difficult for non-experts to understand. In addition, the odds ratio is often misinterpreted as the relative risk, which can lead to erroneous conclusions.

The odds ratio is an important measure of association in observational studies, but it does not provide information about causation. Therefore, it should be interpreted with caution and in conjunction with other statistical measures.

3.3.1. Violence vs Respondent's Education Level

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore

V	0	Experienced no violence
	1	Experienced violence

and E be the random variable denoting the respondent's education level.
Therefore,

E	0	No Education
	1	Primary Education
	2	Secondary Education
	3	Higher Education

The joint pmf of V, E and marginals are:

		E				
		0	1	2	3	P(V = v)
V	0	0.06705	0.03168	0.08066	0.01300	**0.80761**
	1	0.21095	0.11154	0.38237	0.10274	**0.19238**
	P(E = e)	**0.27799**	**0.14321**	**0.46304**	**0.11574**	**1**

The conditional probability distribution of (V = 1 | E = e) is as follows:

P(V = 1 | E = 0)	Probability of Violence among respondents who are not educated	0.24116
P(V = 1 | E = 1)	Probability of Violence among respondents who have primary education	0.22117
P(V = 1 | E = 2)	Probability of Violence among respondents who have secondary education	0.17420
P(V = 1 | E = 3)	Probability of Violence among respondents who have higher education	0.11233

Odds Ratio

Given (V = 1)

(E = 0 V/s E = 1)	1.0904
(E = 0 V/s E = 2)	1.3844
(E = 0 V/s E = 3)	2.1469
(E = 1 V/s E = 2)	1.2696
(E = 1 V/s E = 3)	1.9690
(E = 2 V/s E = 3)	1.5508

Interpretation

- ➢ Respondents who are not educated are 9.04% more likely to experience violence with respect to the respondents who have primary education.
- ➢ Respondents who are not educated are 38.44% more likely to experience violence with respect to the respondents who have secondary education.

- ➢ Respondents who are not educated are 114.69% more likely to experience violence with respect to the respondents who have higher education.
- ➢ Respondents who have primary education are 26.96% more likely to experience violence with respect to the respondents who have secondary education.
- ➢ Respondents who have primary education are 96.90% more likely to experience violence with respect to the respondents who have higher education.
- ➢ Respondents who have secondary education are 55.08% more likely to experience violence with respect to the respondents who have higher education.

3.3.2. Violence vs Respondent's Husband/partner's Education Level

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore,

V	0	Experienced no violence
	1	Experienced violence

and E be the random variable denoting the respondent's Husband/partner's education level. Therefore,

E	0	No Education
	1	Primary Education
	2	Secondary Education
	3	Higher Education

The joint pmf of V, E and marginals are:

		E				
		0	1	2	3	P(V = v)
V	0	0.04153	0.03251	0.09956	0.01877	**0.80760**
	1	0.12772	0.10786	0.442161	0.12985	**0.19239**
	P(E = e)	**0.16925**	**0.140385**	**0.541725**	**0.14863**	**1**

The conditional probability distribution of (V = 1 | E = e) is as follows:

P(V = 1 | E = 0)	Probability of Violence among respondents whose husband/partner are not educated	0.24539
P(V = 1 | E = 1)	Probability of Violence among respondents whose husband/partner have primary degree	0.23163
P(V = 1 | E = 2)	Probability of Violence among respondents whose husband/partner have secondary degree	0.18379
P(V = 1 | E = 3)	Probability of Violence among respondents whose husband/partner have higher degree	0.12633

Odds Ratio

Given (V = 1)

(E = 0 V/s E = 1)	1.0594
(E = 0 V/s E = 2)	1.3352
(E = 0 V/s E = 3)	1.9424
(E = 1 V/s E = 2)	1.2603
(E = 1 V/s E = 3)	1.8335
(E = 2 V/s E = 3)	1.4548

Interpretation

- ➢ Respondents whose husband/partner has no education are 5.94% more likely to experience violence with respect to the respondents whose husband/partner has primary education.
- ➢ Respondents whose husband/partner has no education are 33.52% more likely to experience violence with respect to the respondent's husband/partner who has secondary education.
- ➢ Respondents whose husband/partner are not educated are 94.24% more likely to experience violence with respect to the respondent's husband/partner who has higher education.
- ➢ Respondent’s husband/partner who has primary education is 26.03% more likely to experience violence with respect to the respondent's husband/partner who has secondary education.
- ➢ Respondent’s husband/partner who has primary education is 83.35% more likely to experience violence with respect to the respondent's husband/partner who has higher education.
- ➢ Respondent’s husband/partner who has secondary education is 45.48% more likely to experience violence with respect to the respondent's husband/partner who has higher education.

3.3.3. Violence vs Frequency of Husband/partner being Drunk

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore

V	0	Experienced no violence
	1	Experienced violence

and D be the random variable denoting the frequency of respondent’s husband/partner being drunk.
Therefore,

D	0	Never Drinks
	1	Often
	2	Sometimes

The joint pmf of V, D and marginals are:

		D			
		0	1	2	P(V = v)
V	0	0.07666	0.08682	0.54826	**0.71176**
	1	0.01888	0.04519	0.22416	**0.28823**
	P(D = d)	**0.09554**	**0.13202**	**0.77242**	**1**

The conditional probability distribution of (V = 1 | D = d) is as follows:

P(V = 1 | D = 0)	Probability of Violence among respondents whose partner doesn't drink	0.19761
P(V = 1 | D = 1)	Probability of Violence among respondents whose partner often drinks	0.34233
P(V = 1 | D = 2)	Probability of Violence among respondents whose partner sometimes drinks	0.29020

Odds Ratio

Given (V = 1)

(D = 0 V/s D = 1)	0.5773
(D = 0 V/s D = 2)	0.6810
(D = 1 V/s D = 2)	1.1796

Interpretation

- ➢ Respondents whose partner never drinks are 42.27% less likely to experience violence with respect to the respondents whose partner drinks often.
- ➢ Respondents whose partner never drinks are 31.90% less likely to experience violence with respect to the respondents whose partner drinks sometimes.
- ➢ Respondents whose partner drinks often are 17.96% more likely to experience violence with respect to the respondents whose partner drinks sometimes.

3.3.4. Violence vs Respondent currently working

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore

V	0	Experienced no violence
	1	Experienced violence

and W be the random variable denoting whether the respondent is working or not. Therefore,

W	0	Doesn't Work
	1	Works

The joint pmf of V, W and marginals are:

		W		
		0	1	P(V = v)
V	0	0.06355	0.21680	**0.28036**
	1	0.12883	0.59080	**0.71963**
	P(W = w)	**0.19238**	**0.80761**	**1**

The conditional probability distribution of (V = 1 | W = w) is as follows:

P(V = 1 | W = 0)	Probability of Violence among respondents who don't work	0.66965
P(V = 1 | W = 1)	Probability of Violence among respondents who work	0.73154

Odds Ratio

Given (V = 1)

(W = 0 V/s W = 1)	0.9154

Interpretation

- ➢ Respondents who don't work are 8.46% less likely to experience violence with respect to the respondents who work.

3.3.5. Violence vs Respondent's wealth index

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore,

V	0	Experienced no violence
	1	Experienced violence

and W be the random variable denoting the respondent's wealth index.
Therefore,

W	0	Poor
	1	Middle
	2	Richer
	3	Affluent

The joint pmf of V, W and marginals are:

		W				
		0	1	2	3	P(V = v)
V	0	0.33242	0.16685	0.15887	0.14945	**0.80761**
	1	0.09903	0.04087	0.03244	0.02003	**0.19238**
	P(W = w)	**0.43145**	**0.20773**	**0.19131**	**0.16949**	**1**

The conditional probability distribution of (V = 1 | W = w) is as follows:

P(V = 1 | W = 0)	Probability of Violence among respondent who belong to the poor class	0.22952
P(V = 1 | W = 1)	Probability of Violence among respondent who belong to the middle class	0.19677
P(V = 1 | W = 2)	Probability of Violence among respondent who belong to the rich class	0.16956
P(V = 1 | W = 3)	Probability of Violence among respondent who belong to the affluent class	0.11822

Odds Ratio

Given (V = 1)

(W = 0 V/s W = 1)	1.1664
(W = 0 V/s W = 2)	1.3536
(W = 0 V/s W = 3)	1.9414
(W = 1 V/s W = 2)	1.1605
(W = 1 V/s W = 3)	1.6644
(W = 2 V/s W = 3)	1.4342

Interpretation

- ➢ Respondents who belong to the poor class are 16.64% more likely to experience violence with respect to the respondents who belong to the middle class.

- ➢ Respondents who belong to the poor class are 35.36% more likely to experience violence with respect to the respondents who belong to the rich class.
- ➢ Respondents who belong to the poor class are 94.14% more likely to experience violence with respect to the respondents who belong to the affluent class.
- ➢ Respondents who belong to the middle class are 16.05% more likely to experience violence with respect to the respondents who belong to the rich class.
- ➢ Respondents who belong to the middle class are 66.44% more likely to experience violence with respect to the respondents who belong to the affluent class.
- ➢ Respondents who belong to the rich class are 16.64% more likely to experience violence with respect to the respondents who belong to the affluent class.

3.3.6. Violence vs Respondent's Type of Place of Residence

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore,

V	0	Experienced no violence
	1	Experienced violence

and R be the random variable denoting the respondent's type of place of residence. Therefore,

R	0	Urban
	1	Rural

The joint pmf of V, R and marginals are:

		R		
		0	1	P(V = v)
V	0	0.20668	0.60093	**0.80761**
	1	0.04162	0.15076	**0.19238**
	P(R = r)	**0.24830**	**0.75169**	**1**

The conditional probability distribution of (V = 1 | R = r) is as follows:

P(V = 1 | R = 0)	Probability of Violence among respondent who live in urban area	0.16763
P(V = 1 | R = 1)	Probability of Violence among respondent who live in rural area	0.20056

Odds Ratio

Given (V = 1)

(R = 0 V/s R = 1)	0.8358

Interpretation

- ➢ Respondents who live in urban areas are 16.42% less likely to experience violence with respect to the respondents who live in the rural area.

3.3.7. Violence vs Number of Living Children

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore,

V	0	Experienced no violence
	1	Experienced violence

and C be the random variable denoting the number of living children.
Therefore,

C	0	No Children
	1	1-2 Children
	2	3-5 Children
	3	6+ Children

The joint pmf of V, C and marginals are:

		W				
		0	1	2	3	P(V = v)
V	0	0.07108	0.47111	0.24992	0.01549	**0.80761**
	1	0.01123	0.10290	0.07313	0.00510	**0.19238**
	P(W = w)	**0.08231**	**0.57401**	**0.32305**	**0.02060**	**1**

The conditional probability distribution of (V = 1 | C = c) is as follows:

P(V = 1 | C = 0)	Probability of Violence among those who have no Children	0.13651
P(V = 1 | C = 1)	Probability of Violence among those who have 1-2 Children	0.17926
P(V = 1 | C = 2)	Probability of Violence among those who have 3-5 Children	0.22638
P(V = 1 | C = 3)	Probability of Violence among those who have 6+ Children	0.24801

Odds Ratio

Given (V = 1)

(C = 0 V/s C = 1)	0.7615
(C = 0 V/s C = 2)	0.6030
(C = 0 V/s C = 3)	0.5504
(C = 1 V/s C = 2)	0.7919
(C = 1 V/s C = 3)	0.7228
(C = 2 V/s C = 3)	0.9128

Interpretation

- Respondents who have no children are 23.85% less likely to experience violence with respect to the respondents who have more than or equal to 1 and less than or equal to 2 children.

- ➢ Respondents who have no children are 39.70% less likely to experience violence with respect to the respondents who have more than 2 and less than or equal to 5 children.
- ➢ Respondents who have no children are 44.96% less likely to experience violence with respect to the respondents who have more than 6 children.
- ➢ Respondents who have more than or equal to 1 and less than or equal to 2 children are 20.81% less likely to experience violence with respect to the respondents who have more than 2 and less than or equal to 5 children.
- ➢ Respondents who have more than or equal to 1 and less than or equal to 2 children are 27.72% less likely to experience violence with respect to the respondents who have more than 6 children.
- ➢ Respondents who have more than 2 and less than or equal to 5 children are 8.72% less likely to experience violence with respect to the respondents who have more than 6 children.

3.3.8. Violence vs Pregnancy end in Miscarriage, Abortion, or Stillbirth

Let V be the random variable denoting whether a respondent experienced violence or not.
Therefore,

V	0	Experienced no violence
	1	Experienced violence

and P be the random variable denoting the cause of pregnancy loss.
Therefore,

P	0	Miscarriage
	1	Abortion
	2	Stillbirth

The joint pmf of V, P and marginals are:

		W			
		0	1	2	P(V = v)
V	0	0.51950	0.19764	0.06588	**0.21697**
	1	0.13307	0.06391	0.01999	**0.78302**
	P(P = p)	**0.65257**	**0.26155**	**0.08587**	**1**

The conditional probability distribution of (V = 1 | P = p) is as follows:

P(V = 1 | P = 0)	Probability of Violence among those whose pregnancy ended in miscarriage	0.20391
P(V = 1 | P = 1)	Probability of Violence among those whose pregnancy ended in abortion	0.24436
P(V = 1 | P = 2)	Probability of Violence among those whose pregnancy ended in stillbirth	0.23282

Odds Ratio

Given (V = 1)

(P = 0 V/s P = 1)	0.8345
(P = 0 V/s P = 2)	1.0496
(P = 1 V/s P = 2)	0.8758

Interpretation

- Respondents whose pregnancy ended in miscarriage are 16.55% less likely to experience violence than respondents whose pregnancy ended in abortion.
- Respondents whose pregnancy ended in miscarriage are 4.96% more likely to experience violence than respondents whose pregnancy ended in stillbirth.
- Respondents whose pregnancy ended in abortion are 12.42% less likely to experience violence than respondents whose pregnancy ended in stillbirth.

3.3.9. Violence vs Respondent related to Current Husband prior to Marriage

Let V be the random variable denoting whether a respondent experienced violence or not. Therefore,

V	0	Experienced no violence
	1	Experienced violence

and R be the random variable denoting whether the respondent was related to the current husband prior to marriage.
Therefore,

R	0	No
	1	Yes

The joint pmf of V, R and marginals are:

		R		
		0	1	P(V = v)
V	0	0.02714	0.08359	**0.88925**
	1	0.16523	0.72401	**0.11074**
	P(R = r)	**0.19238**	**0.80761**	**1**

The conditional probability distribution of (V = 1 | R = r) is as follows:

P(V = 1 | R = 0)	Probability of Violence among respondent whose husband was related to them prior to marriage	0.89649
P(V = 1 | R = 1)	Probability of Violence among respondent whose husband was not related to them prior to marriage	0.85888

Odds Ratio

Given (V = 1)

(R = 0 V/s R = 1)	0.9580

Interpretation

- ➢ Respondents whose husband was related to them prior to marriage are 4.20% more likely to experience violence with respect to the respondents whose husband was not related to them prior to marriage.

3.4. Regression Analysis

Regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables. In simpler terms, regression analysis helps to find the relationship between a dependent variable (outcome) and one or more independent variables (predictors). Regression analysis can be used to explore and understand the relationships between variables and to predict the future values of the dependent variable based on the values of the independent variable(s). Regression analysis is often used in many fields, including economics, finance, marketing, biology, psychology, and engineering. It is used to investigate the causal relationship between variables and make predictions based on historical data. There are many types of regression models, including linear regression, logistic regression, polynomial regression, and multiple regression. The choice of model depends on the nature of the data and the research question being addressed.

Linear regression is one of the most widely used regression models. It assumes a linear relationship between the dependent variable and the independent variable(s). The goal of linear regression is to find the best-fit line that explains the relationship between the variables. Multiple regression is used when there are two or more independent variables that affect the dependent variable. The goal of multiple regression is to find the best-fit line that explains the relationship between the variables. Regression analysis is a powerful tool for understanding and predicting the relationships between variables.

A linear model specifies a linear relationship between a dependent variable and independent variables.

$\boldsymbol{y} = \boldsymbol{\beta 0} + \boldsymbol{\beta 1} * \boldsymbol{X1} + \ldots\ldots\ldots\ldots\ldots\ldots + \boldsymbol{\beta n} * \boldsymbol{Xn} + \in$

where y is the dependent variable, {Xi} are independent variables, {βi} are parameters of the model.

3.4.1. Assumptions of Regression Analysis

1. **Linearity**: The relationship between the dependent variable and the independent variable(s) should be linear.
2. **Independence**: The observations should be independent of each other.
3. **Homoscedasticity**: The variance of the errors should be constant across the range of the independent variable(s).
4. **Normality**: The errors should be normally distributed.
5. **No multicollinearity**: The independent variables should not be highly correlated with each other.

3.4.2. Variables

Dependent Variable

1. Proportion of violence in the state (y)

Independent Variables

1. Proportion of respondents whose partner drink in the state (X1)
2. Proportion of respondents who are currently working in the state (X2)
3. Proportion of respondents who live in urban areas in the state (X3)
4. Proportion of respondents who are not educated in the state (X4)
5. Proportion of respondents whose partners are not educated in the state (X5)

3.4.3. Analysis using Google Collab

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

from google.colab import files

uploaded = files.upload()
```

```python
for filename in uploaded.keys():
   print(f"Uploaded file: {filename}")
```

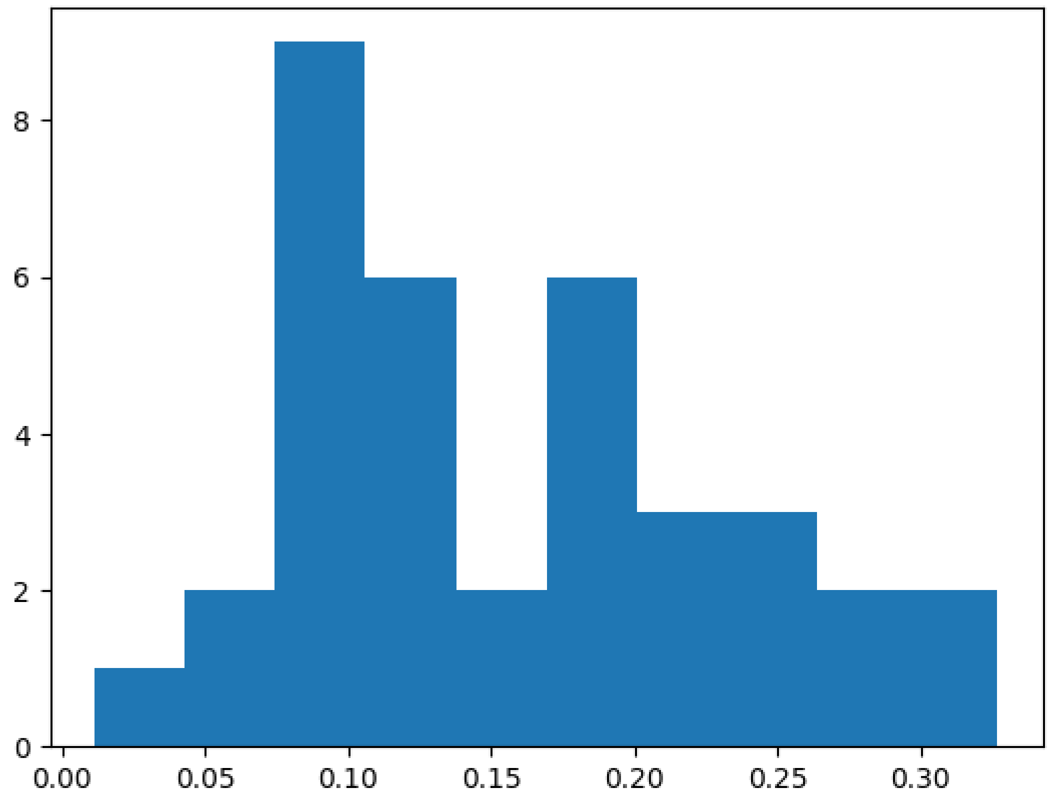
```
Saving Regression.xlsx to Regression.xlsx
Uploaded file: Regression.xlsx
```

```python
df = pd.read_excel('Regression.xlsx')
print(df)
```

```
                                 State         y        X1        X2  \
0                      Jammu & Kashmir  0.082218  0.959677  0.205431
1                     Himachal Pradesh  0.053362  0.774359  0.254478
2                               Punjab  0.076700  0.856566  0.204390
3                           Chandigarh  0.083333  0.785714  0.201550
4                          Uttarakhand  0.104283  0.664062  0.216906
5                              Haryana  0.133014  0.827128  0.186850
6                         Nct Of Delhi  0.141809  0.732026  0.220391
7                            Rajasthan  0.194704  0.822976  0.220339
8                        Uttar Pradesh  0.260906  0.914847  0.164390
9                                Bihar  0.297204  0.927048  0.145197
10                              Sikkim  0.099617  0.991071  0.308300
11                   Arunachal Pradesh  0.131698  0.957950  0.313176
12                            Nagaland  0.080624  0.984496  0.358416
13                             Manipur  0.326923  0.992063  0.400478
14                             Mizoram  0.075410  0.984127  0.256675
15                             Tripura  0.137199  0.926380  0.201407
16                           Meghalaya  0.114146  0.958763  0.392875
17                               Assam  0.223462  0.979472  0.176139
18                         West Bengal  0.162304  0.906818  0.200554
19                           Jharkhand  0.224490  0.968876  0.226693
20                              Odisha  0.231767  0.887786  0.240237
21                        Chhattisgarh  0.173972  0.872795  0.319039
22                      Madhya Pradesh  0.195970  0.904035  0.289160
23                             Gujarat  0.124444  0.929167  0.359604
24   Dadra & Nagar Haveli And Daman & Diu  0.094828  0.923077  0.247059
25                         Maharashtra  0.195030  0.921053  0.378102
26                      Andhra Pradesh  0.194328  0.992021  0.374847
27                           Karnataka  0.289218  0.995720  0.379583
28                                 Goa  0.050314  0.950820  0.297030
29                         Lakshadweep  0.011494  0.988679  0.098266
30                              Kerala  0.082996  0.945946  0.236287
31                          Tamil Nadu  0.277190  0.990000  0.378314
32                          Puducherry  0.237013  0.805825  0.221612
33           Andaman & Nicobar Islands  0.120773  0.990862  0.203085
34                           Telangana  0.256326  0.974359  0.421736
35                              Ladakh  0.198675  0.921321  0.262857

          X3        X4        X5
0   0.819768  0.233016  0.158354
```

```
1    0.916667  0.088542  0.046215
2    0.691930  0.156998  0.161044
3    0.034853  0.120643  0.112360
4    0.805271  0.143524  0.069885
5    0.697613  0.160756  0.098711
6    0.034770  0.125818  0.087809
7    0.787555  0.325820  0.155609
8    0.806162  0.292610  0.192849
9    0.891698  0.392934  0.300659
10   0.841027  0.099052  0.078261
11   0.786947  0.229497  0.220629
12   0.734372  0.100268  0.093122
13   0.723576  0.096120  0.077273
14   0.556395  0.065806  0.055838
15   0.815833  0.122231  0.115668
16   0.877913  0.140882  0.293696
17   0.877326  0.186255  0.199952
18   0.703008  0.198617  0.218904
19   0.805812  0.328326  0.247074
20   0.853098  0.239641  0.184353
21   0.809576  0.276170  0.225870
22   0.789217  0.277381  0.193591
23   0.695198  0.227574  0.139725
24   0.463693  0.159971  0.068404
25   0.668286  0.132425  0.105485
26   0.711982  0.289021  0.340517
27   0.705730  0.209161  0.229173
28   0.389163  0.050739  0.044335
29   0.226904  0.015397  0.032258
30   0.583918  0.010484  0.023865
31   0.575867  0.097583  0.087860
32   0.199237  0.060780  0.106635
33   0.767626  0.081769  0.098246
34   0.727124  0.353623  0.314186
35   0.781316  0.222505  0.138462
```

```
#normality of y
y = df['y']
plt.hist(y, bins=10)
plt.show()
import scipy.stats as st
st.shapiro(y)
```



ShapiroResult(statistic=0.9648677110671997, pvalue=0.3022381067276001)

```
# Calculate the correlation matrix
X = df[['X1', 'X2','X3','X4','X5']]
corr_matrix = X.corr()

# Print the correlation matrix
print(corr_matrix)
```

```
          X1        X2        X3        X4        X5
X1  1.000000  0.350672  0.269499  0.028729  0.231201
X2  0.350672  1.000000  0.182865  0.055613  0.235624
X3  0.269499  0.182865  1.000000  0.473835  0.424134
X4  0.028729  0.055613  0.473835  1.000000  0.814934
X5  0.231201  0.235624  0.424134  0.814934  1.000000
```

```
import numpy as np
from numpy.linalg import inv

def vif_from_corr(corr):
    """
    Calculates the VIF from a correlation matrix
    """
```

```
    # calculate VIF for each predictor variable
    vif = pd.Series(np.linalg.inv(corr_matrix.values).diagonal(),
index=corr_matrix.index)
    vif = 1.0 / vif

    return vif


#Get the vif from correlation matrix
vif = vif_from_corr(corr_matrix)

print(vif)

X1     0.758597
X2     0.820048
X3     0.702675
X4     0.267651
X5     0.275999
dtype: float64

# Apply linear regression model
X = df[['X1', 'X2','X3','X4','X5']]
X = sm.add_constant(X)
y = df['y']
model = sm.OLS(y, X).fit()
print(model.summary())

OLS Regression Results
========================================================================
=======
Dep. Variable:                       y   R-squared:
0.366
Model:                             OLS   Adj. R-squared:
0.261
Method:                  Least Squares   F-statistic:
3.468
Date:                 Wed, 12 Apr 2023   Prob (F-statistic):
0.0136
Time:                         11:45:24   Log-Likelihood:
48.528
No. Observations:                   36   AIC:
-85.06
Df Residuals:                       30   BIC:
-75.55
Df Model:                            5
Covariance Type:             nonrobust
========================================================================
```

```
=======
                 coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------
-------
const         -0.0465      0.138     -0.336      0.739      -0.329
0.236
X1             0.0932      0.162      0.574      0.570      -0.238
0.425
X2             0.2088      0.156      1.338      0.191      -0.110
0.527
X3            -0.0241      0.061     -0.398      0.693      -0.148
0.100
X4             0.4325      0.230      1.884      0.069      -0.036
0.901
X5             0.0377      0.261      0.144      0.886      -0.495
0.571
=======================================================================
=======
Omnibus:                        5.458   Durbin-Watson:
1.874
Prob(Omnibus):                  0.065   Jarque-Bera (JB):
4.565
Skew:                           0.869   Prob(JB):
0.102
Kurtosis:                       3.148   Cond. No.
46.5
=======================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

3.4.4. Interpretation

➢ **Normality of y**

The result of a Shapiro-Wilk test, which is a statistical test we have used to determine if a given sample of data is normally distributed or not.

ShapiroResult(statistic=0.9648677110671997, pvalue=0.3022381067276001)

The "statistic" value (0.9648677110671997) is the test statistic computed by the Shapiro-Wilk test. It measures how far the observed data deviates from what would be expected under the assumption of normality. The closer this value is to 1, the more likely it is that the data is normally distributed.

The ''pvalue' value (0.3022381067276001) is the probability of obtaining a test statistic as extreme as the observed one, assuming that the null hypothesis is true. In this case, the null hypothesis is that the sample comes from a normal distribution. If the p-value is less than a pre-specified significance level (often 0.05), then the null hypothesis is rejected in favor of the alternative hypothesis, which is that the sample is not normally distributed. However, in this case, since the p-value is greater than 0.05, **we fail to reject the null hypothesis and conclude that the sample is normally distributed (at a 5% significance level).**

- **VIF of Variables**

We got the result as:

```
X1    0.758597
X2    0.820048
X3    0.702675
X4    0.267651
X5    0.275999
dtype: float64
```

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can lead to unreliable estimates of the regression coefficients and reduced predictive power of the model.

The VIF values shown here represent the degree of multicollinearity for each variable. The VIF for each variable is a measure of how much the variance of the estimated regression coefficient for that variable is increased because of multicollinearity with the other variables in the model. A VIF value of 1 indicates no multicollinearity, while values greater than 1 indicate the presence of multicollinearity. Generally, a VIF value greater than 5 or 10 is considered high and may require further investigation.

In this case, the VIF values for all variables are less than 1, which suggests that there is little to no multicollinearity among the variables in the regression model. This is a good sign for the model's reliability and predictive power.

➢ **Model Summary**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.366
Model:                            OLS   Adj. R-squared:                  0.261
Method:                 Least Squares   F-statistic:                     3.468
Date:                Wed, 12 Apr 2023   Prob (F-statistic):             0.0136
Time:                        11:45:24   Log-Likelihood:                 48.528
No. Observations:                  36   AIC:                            -85.06
Df Residuals:                      30   BIC:                            -75.55
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0465      0.138     -0.336      0.739      -0.329       0.236
X1             0.0932      0.162      0.574      0.570      -0.238       0.425
X2             0.2088      0.156      1.338      0.191      -0.110       0.527
X3            -0.0241      0.061     -0.398      0.693      -0.148       0.100
X4             0.4325      0.230      1.884      0.069      -0.036       0.901
X5             0.0377      0.261      0.144      0.886      -0.495       0.571
==============================================================================
Omnibus:                        5.458   Durbin-Watson:                   1.874
Prob(Omnibus):                  0.065   Jarque-Bera (JB):                4.565
Skew:                           0.869   Prob(JB):                        0.102
Kurtosis:                       3.148   Cond. No.                         46.5
==============================================================================
```

This output represents the results of a regression analysis for a dependent variable and 5 independent variables.

R-squared (0.366): This is a statistical measure that represents the proportion of variance in the dependent variable that can be explained by the independent variable(s). In this case, 36.6% of the variance in the dependent variable is explained by the independent variable(s).

Adj. R-squared (0.261): This is the adjusted version of R-squared, which takes into account the number of independent variables used in the regression model. It is a better measure of the goodness-of-fit of the model. In this case, the adjusted R-squared is 26.1%.

F-statistic (3.468): This is a statistical test that measures the overall significance of the regression model. It tests the null hypothesis that all regression coefficients are equal to zero. A larger F-statistic indicates a more significant relationship between the dependent variable and the independent variable(s).

Durbin-Watson (1.874): This is a test for autocorrelation in the residuals of the regression model. It measures the degree to which the residuals are correlated with each other. A value of 2 indicates no autocorrelation, while values less than 2 indicate positive autocorrelation and values greater than 2 indicate negative autocorrelation. In this case, the value of 1.874 suggests that there is some positive autocorrelation present in the residuals.

Skew (0.869): This is a measure of the symmetry of the residuals of the regression model. A value of 0 indicates perfect symmetry, while positive values indicate right skew and negative values indicate left skew. In this case, the value of 0.869 suggests that the residuals are slightly right-skewed.

Kurtosis (3.148): This is a measure of the thickness of the tails of the distribution of the residuals of the regression model. A value of 3 indicates a normal distribution, while values greater than 3 indicate heavier tails and values less than 3 indicate lighter tails. In this case, the value of 3.148 suggests that the tails are slightly heavier than normal.

Model :

The full model:

$$y = -0.0465 + 0.0932*X1 + 0.2088*X2 - 0.0241*X3 + 0.4325*X4 + 0.0377*X5 + \varepsilon$$

where:

- y is the dependent variable.
- X1, X2, X3, X4, and X5 are the independent variables.
- ε represents the error term, which captures the variability in y that is not accounted for by the independent variables.

and,

X1 = Proportion of respondents whose partner drinks in the state

X2 = Proportion of respondents who are currently working in the state

X3 = Proportion of respondents who live in urban areas of the state

X4 = Proportion of respondents who are not educated in the state

X5 = Proportion of respondents whose partners are not educated in the state

const (-0.0465): This represents the intercept of the regression model, which is the value of the dependent variable when all of the independent variables are equal to zero.

X1 (0.0932): This coefficient indicates that a one-unit increase in X1 will cause an increase of 0.0932 units in the dependent variable, holding all other variables constant.

X2 (0.2088): This coefficient indicates that a one-unit increase in X2 will cause an increase of 0.2088 units in the dependent variable, holding all other variables constant.

X3 (-0.0241): This coefficient indicates that a one-unit increase in X3 will cause a decrease of -0.0241 units in the dependent variable, holding all other variables constant.

X4 (0.4325): This coefficient indicates that a one-unit increase in X4 will cause an increase of 0.4325 units in the dependent variable, holding all other variables constant.

X5 (0.0377): This coefficient indicates that a one-unit increase in X5 will cause an increase of 0.3777 units in the dependent variable, holding all other variables constant.

Overall, these coefficients show the direction and magnitude of the relationships between the independent variables and the dependent variable in the regression model.

Chapter IV:
Conclusion

Conclusion

- ☆ **Violence v/s Respondent's Education Level**

The respondent's education level has a significant association with violence. It is observed as the education level of the respondents increases it is less likely that she will experience violence.

- ☆ **Violence v/s Respondent's Husband/partner education level**

The respondent's Husband/Partner education level has a significant association with violence. It is observed as the education level of the respondent's husband/partner increases it is less likely that she will experience violence.

- ☆ **Violence v/s Frequency of Husband/partner being drunk**

The respondent's Husband/Partner drinking status has a significant association with violence. It is observed as the frequency of drinking by the respondent's husband/partner increases it is more likely that she will experience violence.

- ☆ **Violence v/s Respondent currently working**

The respondent's working status has a significant association with violence. It is observed that if the respondent works it is more likely that she will experience violence.

- ☆ **Violence v/s Respondent's Wealth Index**

The respondent's wealth index has a significant association with violence. It is observed as the wealth index of the respondent increases it is less likely that she will experience violence.

- ☆ **Violence v/s Type of Place of Residence**

The respondent's place of residence has a significant association with violence. It is observed that the respondent who lives in urban areas are less likely to experience violence.

- ☆ **Violence v/s Number of living children**

The number of children has a significant association with violence. It is observed as the number of children increase it is more likely that she will experience violence.

- ☆ **Violence v/s Pregnancy Loss**

Pregnancy loss has a significant association with violence. It is observed that the respondents whose pregnancy ended in miscarriage are less likely to experience violence with respect to the respondents whose pregnancy ended in abortion but more likely to experience violence with respect to the respondents whose pregnancy ended in stillbirth.

- ☆ **Violence v/s Related to current husband prior to marriage**

Being related to the current husband prior to marriage has a significant association with violence. Respondents whose husband was related to them prior to marriage are slightly more likely to experience violence than the respondents whose husband was not related to them prior to marriage.

- ☆ **Violence v/s Can you say no to your husband if you do not want to have sexual intercourse**

The proportion of violence among the respondents who can say no to intercourse with their husband is higher with respect to those respondents who can't say no to intercourse with their husband.

- ☆ **Violence v/s Caste**

  The proportion of violence among the respondents who belong to the lower caste is higher with respect to those respondents who belong to the upper caste.

Regression Analysis

While fitting a multiple liner regression model between proportion of violence in a state (dependent variable) and various independent variables the very basic assumption i.e. normality of dependent variable was checked using Shapiro-Wilk test, it is found that the dependent variable follows normal distribution.

The presence or absence of multicollinearity was explored by using the concept of VIF(Variance Inflation Factor). It is found that VIF values for all the independent variables are less than 1, which suggest there is little to no multicollinearity among regressors(predictors) incorporated in the regression model.

After the model is fitted, it is explored that:-

- ➢ There is positive impact of proportion of respondents whose partners drink in a state on the proportion of violence in the state i.e. as the proportion of drunkards in a state increases, the proportion of violence in that state will also increase.

- ➢ There is positive impact of proportion of respondents who are currently working in the state on the proportion of violence in the state i.e. as the proportion of working woman in a state increases, the proportion of violence in that state will also increase.

- ➢ Proportion of urban population in a state has a negative impact on proportion of violence in a state i.e. more the urban population leads to less violence.

- ➢ Proportion of women(age between 15 to 49) with no education in the state has a positive impact on proportion of violence in that state, i.e. more the women who are uneducated leads to more violence in a state.

- Proportion of women(age between 15 to 49) whose partners are uneducated in the state has a positive impact on proportion of violence in that state, i.e. more the women whose partners are uneducated leads to more violence in a state.

- It is also observed that only 36.6% of the variance in the dependent variable is explained by the predictors(explanatory variables) in this model, indicating that there are some additional factors or variables that are not included in our study can have an impact on the dependent variable.

References:

1. Fundamentals of Mathematical Statistics: S.C. Gupta , V.K. Kapoor Published by Sultan Chand & Sons, 12th revised edition.
2. https://en.wikipedia.org/wiki/Domestic_violence
3. https://www.un.org/en/coronavirus/what-is-domestic-abuse
4. https://www.undp.org/sustainable-development-goals/gender-equality
5. https://sdgs.un.org/goals/goal5
6. https://dhsprogram.com/
7. https://online.stat.psu.edu/stat415/lesson/9/9.4
8. https://www.bmj.com/about-bmj/resources-readers/publications/statisticssquare-one/8-chi-squared-tests
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/
10. https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.
11. https://towardsdatascience.com/a-practical-introduction-to-the-shapiro-wilk-test-for-normality-5675e52cee8f
12. https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factor%20measures%20how,standard%20error%20in%20the%20regression.

THANK YOU