# Violence Against Women and Girls

## BootCamp on Data Science and Tools

Submitted

To

## CRC-Training


**By**

**Ritika Baghel**

**Soumya Agarwal**


**Under the guidance of**
**Mr. Gaurav Kansal / Mr. Gopal Gupta**

# TABLE OF CONTENTS

# CERTIFICATE

This is to certify that the Project Report entitled **"Violence Against Women and Girls"** which is submitted by Ritika Baghel, and Soumya Agarwal and in partial fulfilment of the requirement for the "Bootcamp on Data Science and Tools" in the Department of CRC-Training of ABES Institute of Technology, is a record of the candidate own work carried out by him under my/our supervision.

**Mr. Gaurav Kansal**

**Mr. Gopal Gupta**

**Date:**

# ABSTRACT

This dataset aggregates agreement with key questions across gender, education level and many other socioeconomic variables from 70 different countries.

The data was collected as part of the Demographic and Health Surveys (DHS) program, which exists to advance the global understanding of health and population trends in developing countries.

The project is based on how much percentage of individual is agreed on key question across, educational level, or many other socioeconomic variables.

The Dataset we have worked upon is downloaded from Kaggle where we have the csv file. Firstly, we analyzed the data set by simply opening the “csv” file. Then we pre-processed the whole data set to remove the insignificant columns and empty rows. Also, to get to know all unique values of each attribute.

Further, we performed the data visualization on the data set. We performed Exploratory data analysis (EDA Analysis) of each to get an idea about how the values of different columns are related to each other.

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Problem Definition: -

- While doing analysis on the project we are analyzing which country people have responded the most against violence.
- Therefore, to make it easier we are using data analysis to analyze the percentage of violence.
- We are analyzing 70 countries in which we see which category (education, residence, marital status, gender) of people give maximum responses over that the husband is justified to hitting his wife.

## 1.2 Motivation:

- As already stated in our problem statement, this model helps in predicting violence against women.
- In the current situation the world is hit by about 1 in 3(30%) of women worldwide have been subjected to either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime estimated by WHO.
- By using the data analysis, we will be predicting the rate of violence in a particular country.

**1.3 Objective of the Project:** The Project is based on a data analysis, that can analyze the responses to violence of 70 countries over some different questions.

# CHAPTER 2

## RELATED WORK

We have read many articles on websites about our projects. We got much useful information from it about data sets and attributes in our data sets. This section will contain references to two such papers. After reading the papers we got these basic ideas about datasets:

- The first paper is of THE LANCET, which is written by Mary Ellsberg, Diana J Arango, Matthew Morton, Floriza Gennari,Sveinung Kiplesund, Manuel Contreras, Charlotte Watt Under the supervision of Markus Hofmann.
- This project will investigate the data mining of demographic data in order to create one or more classification models which are capable of accurately identifying individuals whose salary exceeds a specified value.
- The second paper is written by Charlotte Bunch on the intolerable status quo: Violence against women and girls.

- According to WHO study published in 2021 that estimated global and regional prevalence of intimate partner violence and non-partner sexual violence against women.
- This report is based on an analysis of available prevalence data from surveys and studies conducted between 2000 and 2018, obtained through a systematic and comprehensive review of all available data on the prevalence of these two forms of violence against women.
- The papers described that the data set belongs to the United States of America, “since the start of COVID-19, 45 per cent of women reported that they or a woman they know has experienced a form of VAWG (Violence Against Women and Girls)”, according to the statement.
- Globally 81,000 women and girls were killed in 2020, around 47,000 of them (58 per cent) died at the hands of an intimate partner or a family member.
- Construct data set for categorical values.

# CHAPTER 3

## PROPOSED METHODOLOGY

### 3.1 Dataset Description

- The dataset for our Analysis was downloaded from Kaggle.
- The data set belongs to 70 Countries.
- The data set has 12600 rows.
- It has a total of 8 attributes. They are as follows:

**a. Record id:** -Numeric value unique to each question by country.

**b. Gender:** - Whether the respondents was Male or Female.

**c. Demographics Question: -** Refers to the different types of demographic groupings used to segment respondents – marital status, education level, employment status, residence type, or age.

**d. Demographics Response: -** Refers to the demographic segment into which the respondent falls (e.g., the age groupings are split into 15-24, 25-34, and 35-49)

- Age (15-24, 25-34, 35-49)
- Education (No Education, Primary, Secondary, Higher)
- Employment (Unemployed, employed for cash, Employed for kind)
- Marital status (Married or living together, Widowed, divorced, separated, Never Married)
- Residence (Rural, Urban)

**e. Marital Status: -** Respondents were asked if they agreed with the following statements:

- husband is justified in hitting or beating his wife if she burns the food.
- husband is justified in hitting or beating his wife if she argues with him.
- husband is justified in hitting or beating his wife if she goes out without telling him.
- husband is justified in hitting or beating his wife if she neglects the children.
- husband is justified in hitting or beating his wife if she refuses to have sex with him.
- husband is justified in hitting or beating his wife for at least one specific reason.

**f. Survey Year: -** Year in which the Demographic and Health Survey (DHS) took place.

**g. Value: -** Value % of people surveyed in the relevant group who agree with the question e.g., the percentage of women aged 15-24 in Afghanistan who agree that a husband is justified in hitting or beating his wife if she burns the food

### 3.2 Methods:

After downloading the data set, we begin by analyzing the data set's attributes, and then preprocessing it.
With the help of **matplotlib** and **Seaborn** we visualize each attribute with a target variable.

## PANDAS:

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Panda is an open-source library that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series.

The first step of working in pandas is to ensure whether it is installed in the Python folder or not. If not, then we need to install it in our system using **the pip command**. Type cmd command in the search box and locate the folder using cd command where **python-pip file** has been installed. After locating it, type the command:

```
pip install pandas
```

After the pandas have been installed into the system, you need to import the library. This module is generally imported as:

```
In [2]: #for manipulation
        import pandas as pd
```

Here, pd is referred to as an alias to the Pandas.

**Advantages:**

- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN).
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.

**Pandas Series:**

Series is a one-dimensional array that can contain any type of data, you can create a series by using the following constructor.

pandas.Series(data, index, dtype, copy)


**Pandas Dataframe:**

A DataFrame is a multi-dimensional data structure in which data is arranged in the form of rows and columns. You can create a DataFrame using the following constructor:

pandas.DataFrame(data, index, columns, dtype, copy)

**NUMPY:**

NumPy stands for numeric python which is a python package for the computation and processing of the multidimensional and single dimensional array elements.

It provides various functions which are capable of performing the numeric computations with a high speed.

NumPy provides various powerful data structures, implementing multi-dimensional arrays and matrices. These data structures are used for the optimal computations regarding arrays and matric

Install numpy by using pip command in command prompt

```
pip install numpy
```

Import numpy in jupyter notebook

```
import numpy as np
```

**Advantages:**

NumPy arrays are faster and more compact than Python lists.

An array consumes less memory and is convenient to use.

NumPy uses much less memory to store data and it provides a mechanism of specifying the data types. This allows the code to be optimized even further.

## MATPLOTLIB:

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc.

It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.

the package is imported into the Python script by adding the following statement −

```
#for visualization
import matplotlib.pyplot as plt
```

**Advantages:**

Matplotlib supports various types of graphical representations like Bar Graphs, Histograms, Line Graph, Scatter Plot, Stem Plots, etc.

Matplotlib can be used in multiple ways including Python scripts, the Python and iPython shells, Jupyter Notebooks.

Matplotlib is a 2-D plotting library. But there are some extensions that we can use to create advanced visualizations like 3-Dimensional plots, etc.

## SEABORN:

The Python library Seaborn is a data visualization library based on Matplotlib.

Seaborn provides a variety of visualization patterns. It is more integrated to work with Pandas dataframe compared to matplotlib. Seaborn is widely used for statistics visualization because it has some of the best statistical tasks built with-in.

**Advantages:**

Seaborn uses fewer Syntax and we write less code to achieve high-grade visualizations.

When compared to matplolib, the seaborn graphs are much more visually appealing by default.

Seaborn works with the whole dataset as a whole compared to matplotlib which deals with dataframes and arrays.

## MATPLTLOB vs SEABORN

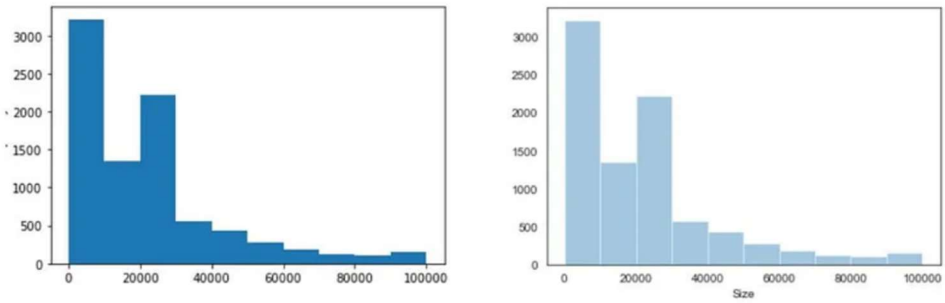


Fig.3.1 Difference between Matplotlib and Seaborn

## IPY Widgets:

Widgets in ipython are GUI based interaction tools provided within the ipython interpreter console. It helps to interact with different components by real-time changing the value of integers based on the widget used. To install it, use the following command in jupyter notebook. The ipywidgets are modules in python to use widget within the jupyter cells. There are many types of widgets provided under this liberality. In ML most of the time it is used to understand the importance of the features within the model and thus choose only the best ones.

The ipywidgets library provides a useful collection of interactive widgets (sliders, check boxes, radio buttons, etc.) for use in the Jupyter Notebook and in several other contexts . For the full list of built-in widgets see . The integration of graphical widgets into the notebook workflow allows users to configure ad-hoc control panels to interactively sweep over parameters using graphical widget controls, rather than by editing code or writing loops over fixed ranges of value.

### 3.3 Software Requirements:

Pandas: 1.1.5
Numpy; 1.19.5
Matplotlib: 3.1.0
Seaborn: 0.11.2
Python: 3.7

### 3.4 Our Methodology:

- Firstly, we downloaded the data set from Kaggle for analysis.
- Then we loaded the data set into Jupyter Notebook

```
In [1]: #for manipulation
        import pandas as pd
        import numpy as np
        #for visualization
        import matplotlib.pyplot as plt
        import seaborn as sns
        from ipywidgets import interact
```

- 

**Fig. 3.2 Importing numpy and pandas and other libraries.**

Importing numpy, pandas, matplotlib, and seaborn packages

NumPy is a Python library used for working with arrays.

Pandas is a high-level data manipulation tool.

Matplotlib is a Python library for creating static, animated and interactive visualizations in Python.

Seaborn is a Python data visualization library based on matplotlib.

```
In [4]: #Lets read the dataset
        data=pd.read_csv('D:\downloads\dataset (2)\Violence Against Women  Girls Data.csv')
```

**Fig. 3.3: read_csv()**

```
In [84]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12600 entries, 0 to 12599
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   RecordID               12600 non-null  int64
 1   Country                12600 non-null  object
 2   Gender                 12600 non-null  object
 3   Demographics Question  12600 non-null  object
 4   Demographics Response  12600 non-null  object
 5   Question               12600 non-null  object
 6   Survey Year            12600 non-null  object
 7   Value                  11187 non-null  float64
dtypes: float64(1), int64(1), object(6)
memory usage: 787.6+ KB
```

**Fig. 3.4: info() function**

The **info()** function is used to information about each and every attribute.

- Total there are 8 columns.
- There are 6 object variable
- There are 1 int64 variable
- There are 1 float64 variable.
- There are 12600 rows entries

The total size of data set is 787.6+KB

```
In [85]: data.head()
```

Out[85]:

| | RecordID | Country | Gender | Demographics Question | Demographics Response | Question | Survey Year | Value |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Afghanistan | F | Marital status | Never married | ... if she burns the food | 01/01/2015 | NaN |
| 1 | 1 | Afghanistan | F | Education | Higher | ... if she burns the food | 01/01/2015 | 10.1 |
| 2 | 1 | Afghanistan | F | Education | Secondary | ... if she burns the food | 01/01/2015 | 13.7 |
| 3 | 1 | Afghanistan | F | Education | Primary | ... if she burns the food | 01/01/2015 | 13.8 |
| 4 | 1 | Afghanistan | F | Marital status | Widowed, divorced, separated | ... if she burns the food | 01/01/2015 | 13.8 |

**Fig. 3.5: head() function**

Using the **head() function** to get a basic idea about how the data is entered.

## ➢ DATA CLEANING: -

```
In [109]: dn.duplicated()

Out[109]: 1        False
          2        False
          3        False
          4        False
          5        False
                   ...
          12595    False
          12596    False
          12597    False
          12598    False
          12599    False
          Length: 11187, dtype: bool
```

**Fig. 3.6: check Duplicate values.**

- There are 12600 rows which does not have any duplicate values.

## ➢ CHECKING MISSING VALUES FOR EACH COLUMN

We know that the missing values in our dataset are represented as "?"

```
In [24]: data.isnull().sum()

Out[24]: RecordID                    0
         Country                     0
         Gender                      0
         Demographics Question       0
         Demographics Response       0
         Question                    0
         Survey Year                 0
         Value                    1413
         dtype: int64
```

**Fig. 3.6: isnull().sum() function**

After running this, we get that
there are: -

- **"1413"** null values are there in Values attribute.

So, we have to check these values.

```
In [27]: # lets check if there is any missing value in the dataset
         data.isna()
```

Out[27]:

| | RecordID | Country | Gender | Demographics Question | Demographics Response | Question | Survey Year | Value |
|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | True |
| 1 | False | False | False | False | False | False | False | False |
| 2 | False | False | False | False | False | False | False | False |
| 3 | False | False | False | False | False | False | False | False |
| 4 | False | False | False | False | False | False | False | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 12595 | False | False | False | False | False | False | False | False |
| 12596 | False | False | False | False | False | False | False | False |
| 12597 | False | False | False | False | False | False | False | False |
| 12598 | False | False | False | False | False | False | False | False |
| 12599 | False | False | False | False | False | False | False | False |

12600 rows × 8 columns

**Fig. 3.7: isna() function.**

## ➢ REMOVE ROWS WITH MISSING VALUES

- ✓ Then using **dropna()** function we remove all Rows with NaN values.
- ✓ It can be seen that total row is reduced to 11187 from 12600.

```
In [26]: #DROP ROWs
         dn=data.dropna()
         dn
```

Out[26]:

| | RecordID | Country | Gender | Demographics Question | Demographics Response | Question | Survey Year | Value |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Afghanistan | F | Education | Higher | ... if she burns the food | 01/01/2015 | 10.1 |
| 2 | 1 | Afghanistan | F | Education | Secondary | ... if she burns the food | 01/01/2015 | 13.7 |
| 3 | 1 | Afghanistan | F | Education | Primary | ... if she burns the food | 01/01/2015 | 13.8 |
| 4 | 1 | Afghanistan | F | Marital status | Widowed, divorced, separated | ... if she burns the food | 01/01/2015 | 13.8 |
| 5 | 1 | Afghanistan | F | Employment | Employed for kind | ... if she burns the food | 01/01/2015 | 17.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 12595 | 210 | Zimbabwe | M | Residence | Urban | ... if she goes out without telling him | 01/01/2015 | 11.8 |
| 12596 | 280 | Zimbabwe | M | Residence | Rural | ... if she neglects the children | 01/01/2015 | 20.1 |
| 12597 | 280 | Zimbabwe | M | Residence | Urban | ... if she neglects the children | 01/01/2015 | 15.0 |
| 12598 | 350 | Zimbabwe | M | Residence | Rural | ... if she refuses to have sex with him | 01/01/2015 | 7.2 |
| 12599 | 350 | Zimbabwe | M | Residence | Urban | ... if she refuses to have sex with him | 01/01/2015 | 3.7 |

11187 rows × 8 columns

**Fig. 3.8: Removing rows with null values.**

## ➢ CHECK THAT ALL NULL VALUES ARE REMOVE:

Here we use again isna() function to check null values.

```
In [28]: dn.isnull().sum()

Out[28]: RecordID                0
         Country                 0
         Gender                  0
         Demographics Question   0
         Demographics Response   0
         Question                0
         Survey Year             0
         Value                   0
         dtype: int64
```

**Fig. 3.1: No null values exist**

- **DATA VISUALIZATION**

## COUNTRY

```
In [30]: country_dn=dn.groupby(["Country"])["Value"].agg(["median","max","min","mean"]).reset_index()
         country_dn.sort_values(['Country','max'])
```

Out[30]:

| | Country | median | max | min | mean |
|---|---|---|---|---|---|
| 0 | Afghanistan | 44.50 | 86.9 | 4.5 | 42.683333 |
| 1 | Albania | 3.75 | 29.7 | 0.0 | 5.417778 |
| 2 | Angola | 11.80 | 37.4 | 0.6 | 13.202222 |
| 3 | Armenia | 6.95 | 47.0 | 0.0 | 9.553571 |
| 4 | Azerbaijan | 33.95 | 74.3 | 1.5 | 32.666071 |
| ... | ... | ... | ... | ... | ... |
| 65 | Uganda | 23.60 | 59.8 | 1.5 | 24.840556 |
| 66 | Ukraine | 2.35 | 22.0 | 0.1 | 3.692308 |
| 67 | Yemen | 28.20 | 61.6 | 1.3 | 27.621111 |
| 68 | Zambia | 23.90 | 63.7 | 0.7 | 24.361667 |
| 69 | Zimbabwe | 16.60 | 55.2 | 1.1 | 18.170556 |

70 rows × 5 columns

**Fig. 3.2: MEAN, MEDIAN OF 70 COUNTRIES**

```
In [155]: sns.catplot(
           data=dn, y="Country", x="Value",
           kind='bar', height=12,aspect=.6,
          )
```

```
Out[155]: <seaborn.axisgrid.FacetGrid at 0x1cb3c3d6e48>
```

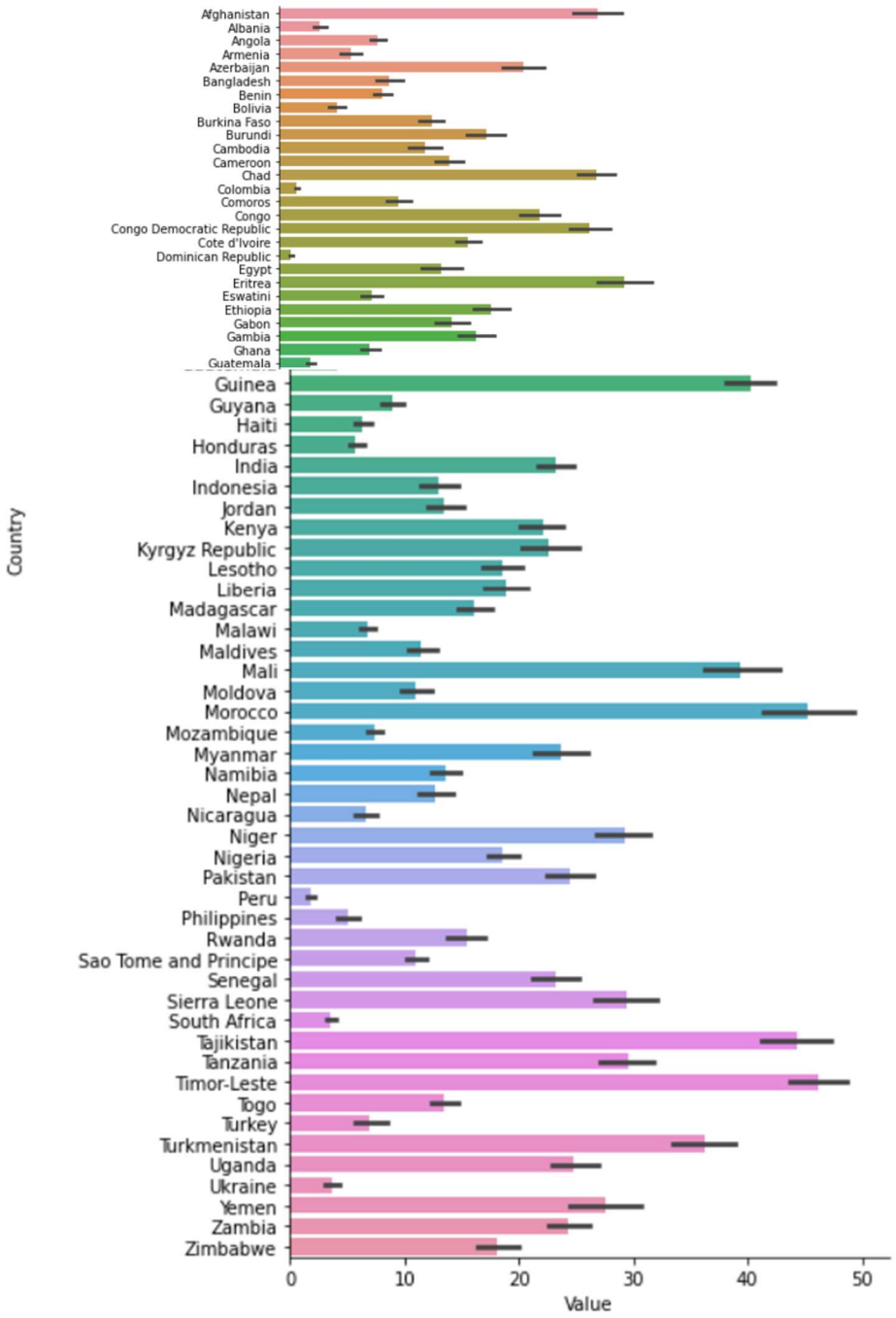


**Fig. 3.3: catplot of country**

- Above Visualization, shows the Response distribution among the Countries in our dataset.
- Here, Eritrea is the country in which the people give maximum responses around the Questions.

## GENDER

```
In [44]: g = sns.catplot(x='Gender',y='Value',
                         data=dn,kind='bar',errorbar=None,aspect=1)
         g.set_axis_labels('Gender','Percentage (%)')
         g.fig.suptitle(' a husband is justified in hitting his wife',y=1.05)

Out[44]: Text(0.5, 1.05, ' a husband is justified in hitting his wife')
```

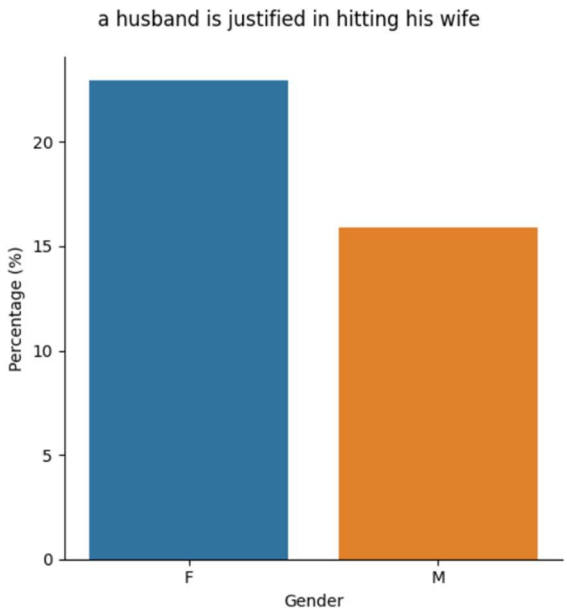


**Fig. 3.4: Gender vs. Percentage**

- Looking at the graph, we can see that the percentage of Female is greater than Male.
- That means, that the Female have the maximum responses over the question that "Husband is justified for hitting his wife"

## Questions Asked

```
In [37]: g = sns.catplot(x='Question',y='Value',data=dn,kind='bar',errorbar=None, height=4,
             aspect=1,)
         g.set_axis_labels('Questions Asked','Percentage (%)',)
         g.fig.suptitle('A husband is justified in hitting his wife',y=1.05)
         plt.xticks(rotation=90);
```

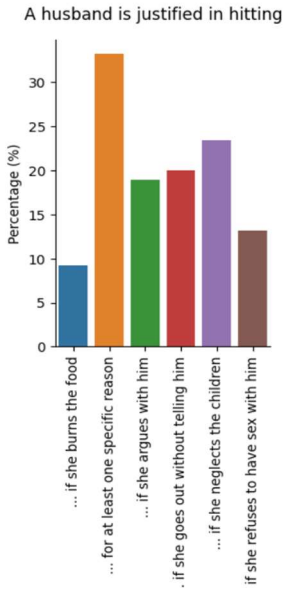


**Fig. 3.5: Questions vs. Percentage**

From the Graph of Questions and Percentage.

- We can see that the there is maximum number of people were agreed on that the husband is justified to hitting his wife “for at least one specific reason”.
- And minimum is agreed on the question that “if she burns the food”.

## Residence

```
In [41]: g = sns.catplot(x='Demographics Response',y='Value',
                         order=['Rural','Urban'],
                         data=dn,kind='bar',errorbar=None)
         g.set_axis_labels('Residence','Percentage (%)')
         g.fig.suptitle('A husband is justified in hitting his wife',y=1.05)

Out[41]: Text(0.5, 1.05, 'A husband is justified in hitting his wife')
```

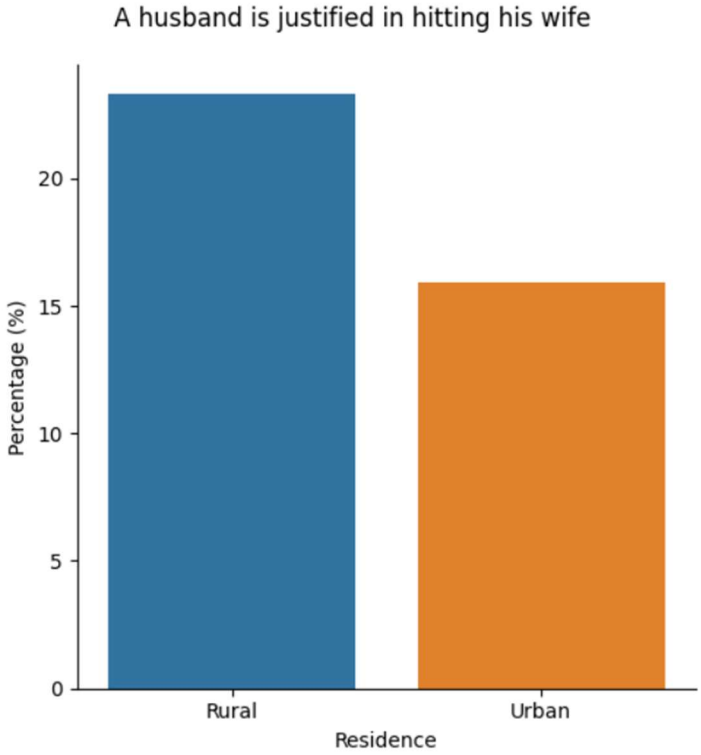


**Fig. 3.6: Residence vs. Percentage**

From the Graph of Residence and Percentage

- We can see that, the Rural peoples are more respond on that “the husband is justified to hitting his wife”

**AGE:**

```
In [42]: g = sns.catplot(x='Demographics Response',y='Value',
                         order=['15-24','25-34','35-49'],
                         data=dn,kind='bar',errorbar=None)
         g.set_axis_labels('Age','Percentage (%)')
         g.fig.suptitle('A husband is justified in hitting his wife',y=1.05)

Out[42]: Text(0.5, 1.05, 'A husband is justified in hitting his wife')
```

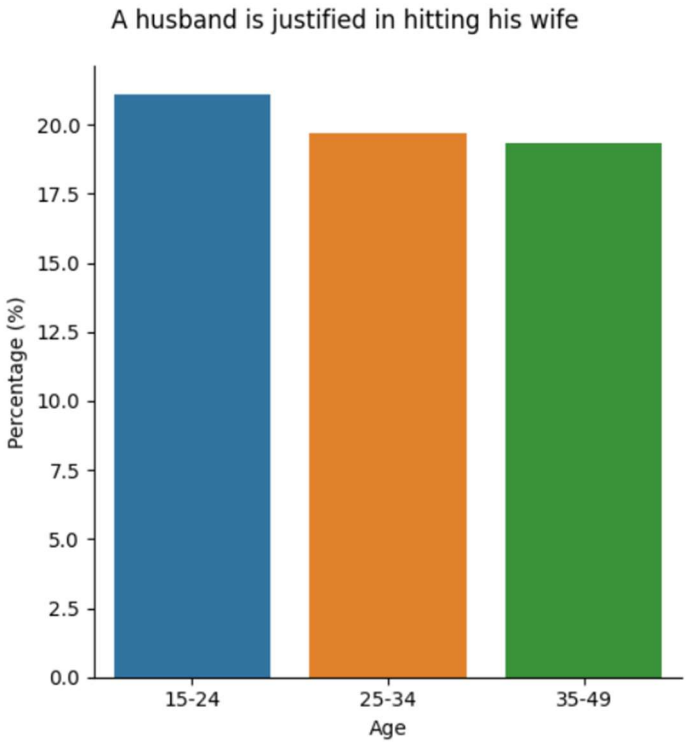


**Fig. 3.7: Age vs. Percentage**

From the Graph of Age and Percentage.

- Under the age of (15-24) these ages of people give the maximum responses that the “husband is justified for hitting his wife”

**EDUCATION:**

```
In [44]: g = sns.catplot(x='Demographics Response',y='Value',
                         order=['No education','Primary','Secondary','Higher'],
                         data=dn,kind='bar',errorbar=None)
         g.set_axis_labels('Education Level','Percentage (%)')
         g.fig.suptitle('A husband is justified in hitting his wife',y=1.05)

Out[44]: Text(0.5, 1.05, 'A husband is justified in hitting his wife')
```

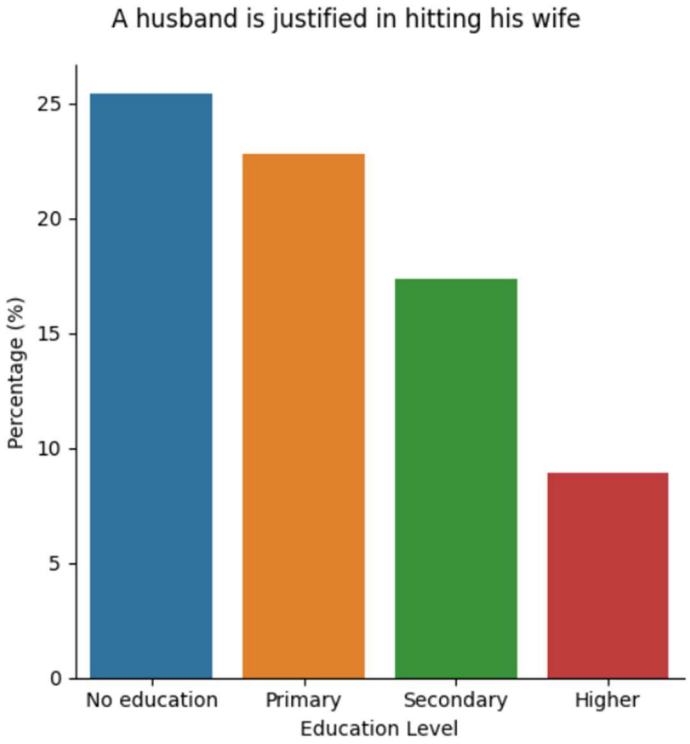


**Fig. 3.8: Education vs. Percentage**

From the Graph of Education and Percentage

- Here, we can see that the the people with No education gave the maximum responses.
- And the people which have the Higher education give the minimum responses.

## Employment

```
In [163]: g = sns.catplot(x='Demographics Response',y='Value',
           order=['Employed for cash','Unemployed','Employed for kind'],aspect=.9,
           data=dn,kind='bar')
          g.set_axis_labels('Employment','Percentage (%)')
          g.fig.suptitle('Indian agreeing a husband is justified in hitting his wife',y=1.05)

Out[163]: Text(0.5, 1.05, 'Indian agreeing a husband is justified in hitting his wife')
```

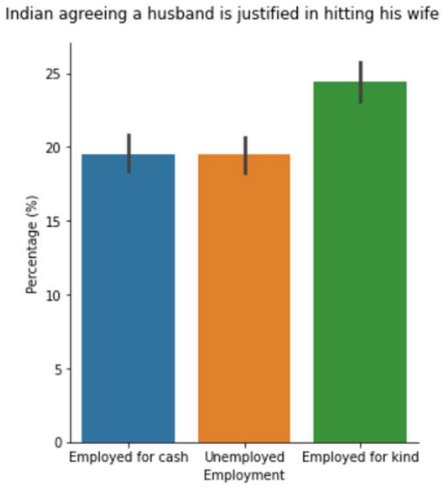


**Fig. 3.9: Employment vs. Percentage**

From the Graph of Employment and Percentage.

- Here, we can see that employed for kind gave maximum number of responses.
- Disregarding employed for cash and unemployed.

10

## MARITAL STATUS

From the Graph of Marital Status and Total Entries.

- We can clearly see that number of Married-civ-spouse has the highest entry with single as second.
- One concerning thing is divorced entry is at third highest point

```
In [47]: g = sns.catplot(x='Demographics Response',y='Value',
                         order=['Never married','Married or living together','Widowed, divorced, separated'],
                         data=dn,kind='bar',errorbar=None)
g.set_axis_labels('Mariage Status','Percentage (%)')
g.fig.suptitle('A husband is justified in hitting his wife',y=1.05)
plt.xticks(rotation=20);
```

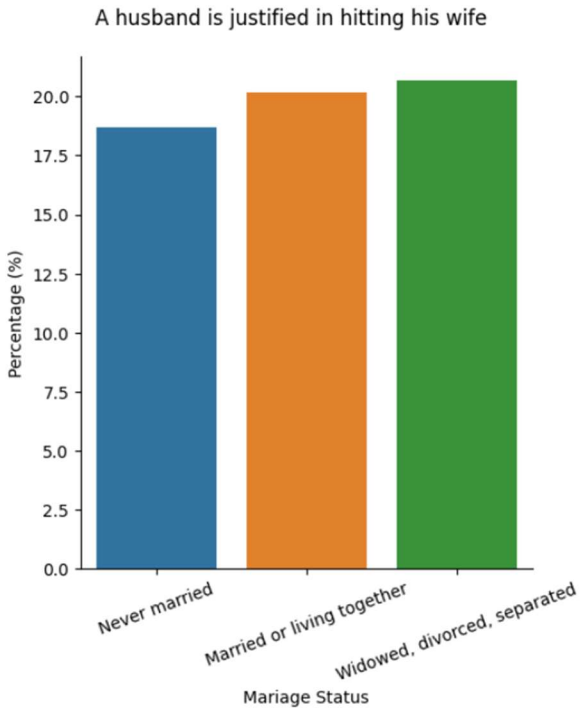


**Fig. 3.11: Marital status vs. No. of entries**



**Survey Year:**

```
In [176]: sns.lineplot(data=dn, x="year", y="Value",hue="Gender")
          plt.xticks(rotation=10);
```

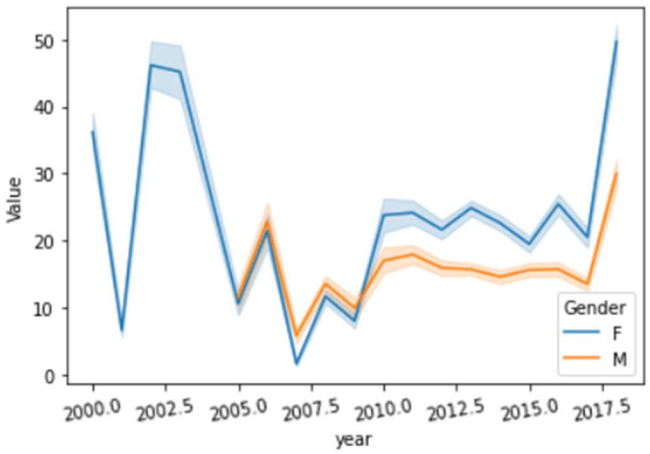


**Fig. 3.13: Survey year vs. Value**

From the Graph of Survey year and Value.

We can see that the percentage of female are very large in giving responses than others in the year 2000 to 2018.

14151617181920.21222324

After this we take India for data visualzation

**INDIA:**

INDIA

```
In [48]: India = dn[dn.Country == 'India']
         India.head()
```

Out[48]:

| | RecordID | Country | Gender | Demographics Question | Demographics Response | Question | Survey Year | Value |
|---|---|---|---|---|---|---|---|---|
| **5580** | 32 | India | F | Education | Higher | ... if she burns the food | 01/01/2015 | 11.4 |
| **5581** | 32 | India | F | Residence | Urban | ... if she burns the food | 01/01/2015 | 15.2 |
| **5582** | 32 | India | F | Marital status | Never married | ... if she burns the food | 01/01/2015 | 15.8 |
| **5583** | 32 | India | F | Education | Secondary | ... if she burns the food | 01/01/2015 | 17.3 |
| **5584** | 32 | India | F | Age | 15-24 | ... if she burns the food | 01/01/2015 | 17.5 |

**Fig. 3.25: India head() function**

**Education:**

```
In [47]: g = sns.catplot(x='Demographics Response',y='Value',col='Gender',hue='Question',
                         order=['No education','Primary','Secondary','Higher'],
                         data=dn,kind='bar',errorbar=None)
         g.set_axis_labels('Education Level','Percentage (%)')
         g.fig.suptitle('India agreeing a husband is justified in hitting his wife',y=1.05)
```

Out[47]: Text(0.5, 1.05, 'India agreeing a husband is justified in hitting his wife')

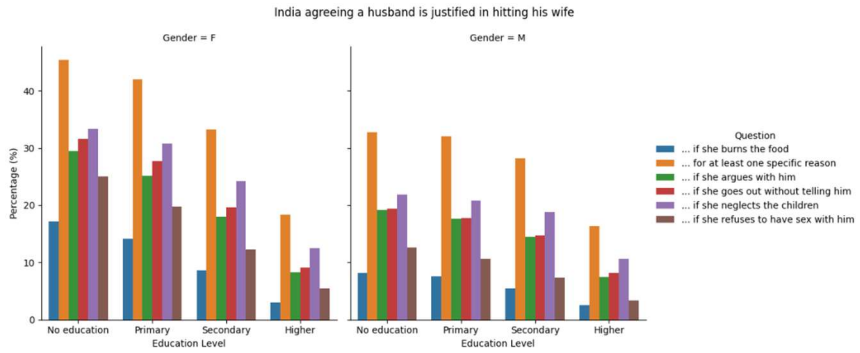


**Fig. 3.26:India Education level VS Percentage**

**Employment:**

```
In [54]: graph = India[(India['Demographics Question'] == 'Employment')]

         g = sns.catplot(x='Demographics Response',y='Value',col="Gender",
                         order=['Employed for cash','Unemployed'],aspect=.9,
                         data=graph,kind='bar',errorbar=None)
         g.set_axis_labels('Employment','Percentage (%)')
         g.fig.suptitle('Indian agreeing a husband is justified in hitting his wife',y=1.05)

Out[54]: Text(0.5, 1.05, 'Indian agreeing a husband is justified in hitting his wife')
```


**Fig. 3.27:India Employment level VS Percentage**

## Residence:

```
In [55]: graph = India[(India['Demographics Question'] == 'Residence')]

         g = sns.catplot(x='Demographics Response',y='Value',col="Gender",
                         order=['Rural','Urban'],aspect=.9,
                         data=graph,kind='bar',errorbar=None)
         g.set_axis_labels('Residence','Percentage (%)')
         g.fig.suptitle('Indian agreeing a husband is justified in hitting his wife',y=1.05)

Out[55]: Text(0.5, 1.05, 'Indian agreeing a husband is justified in hitting his wife')
```

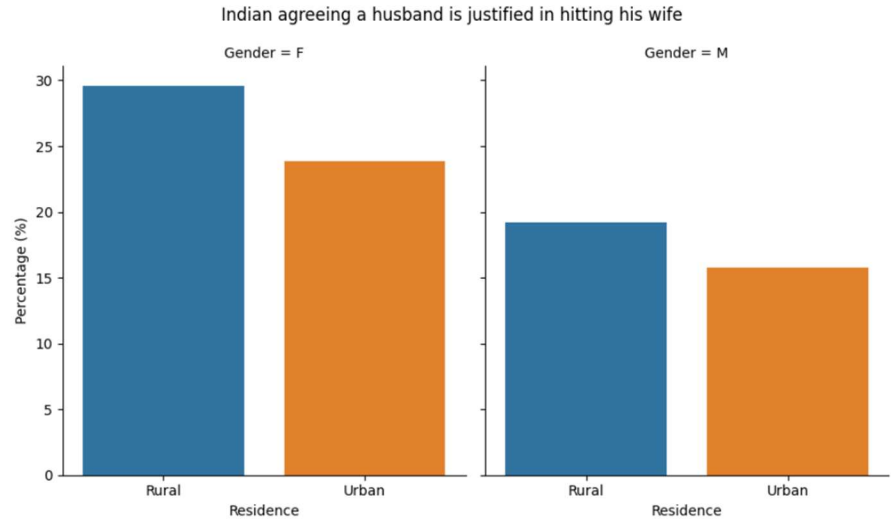


**Fig. 3.28: India Residence vs Percentage**

# CHAPTER 4

## RESULT ANALYSIS

a)2018 was the year that recorded with the most agreed key questions that were asked.

b) The country which had the highest % average of people who agreed with the key question was Eritrea with an average of 46.19%.

c) The most agreeable question was that the husband is justified in beating his wife for at least one specific reason with the highest % average being in 2018 for females at 70.39% and males at 48.84%.

d) We can also see that females have the highest % of responses compared to males across all key question this could be due to some socioeconomic factors affecting the communities across the world.

e) In the demographic age we can also see that ages 15-24 has the highest % average of agreement with the key question husband is justified in beating his wife for at least one specific reason coming to 35.88% which is the highest across other age groups.

f) In the demographic Education group we can see that for females those who have received no education agree to the key question that a husband is justified to beating his wife for at least one specific reason with a percentage of 45.45% compares to males in the same education category where there average % was 32.81%.

g) In the demographic group of residence we see that those who live in rural areas have a highest % responses of the key question that husband is justified in beating his wife for at least one specific reason coming to 42.24% for both genders and the ones in urban at 30.94%.

# CHAPTER 5
# CONCLUSION

## 5.1 Discussion

After completing my first project on Data Analysis and Machine Learning, we concluded that our data set gave us the best accuracy on the key question that a husband is justified in hitting his wife for any reason has the highest % of agreed responses from both male and female, we can conclude this could be because of the social construct that a man is the head of the house so they do not need to explain themselves when it comes to abuse or violence in their homes, the way forward is to start teaching equality in homes that a wife is as important in the household as a husband, changing the view of what is commonly known as a household hierarchy.

## 5.2 Future Work

- Future work on our project is as follows:
- Make the machine learning model.
- Make the GUI of the project

# REFERENCES


## Downloading data set:

https://www.kaggle.com/datasets/andrewmvd/violence-against-women-and-girls

Demographic and Health Surveys (DHS)

## Loading and analyzing

https://pandas.pydata.org/

https://numpy.org/

## Pre-processing

https://towardsdatascience.com/data-preprocessing-in-python-b52b652e37d5

https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/

https://scikit-learn.org/stable/modules/preprocessing.html

https://www.geeksforgeeks.org/data-cleansing-introduction/

## Data Visualization

https://matplotlib.org/tutorials/index.html

https://python-graph-gallery.com/seaborn/