SURVEY ON WOMEN PARTICIPATION IN NIGERIA POLITICS

Problem Statement

Importing necessary libraries

In [1]:
```
# Installing the libraries with the specified version.
# uncomment and run the following line if Google Colab is being used
# !pip install scikit-learn==1.2.2 seaborn==0.13.1 matplotlib==3.7.1 numpy==1.25.2 pandas==1.5.3 imbalanced-learn==0.10.
```

In [1]:
```
# Installing the libraries with the specified version.
# uncomment and run the following lines if Jupyter Notebook is being used
# !pip install scikit-learn
# !pip install seaborn
# !pip install matplotlib
# !pip install numpy
# !pip install pandas
# !pip install imblearn
# # !pip install xgboost -q --user
# # !pip install --upgrade -q threadpoolctl
# !pip install scikit-plot
```

```
Requirement already satisfied: scikit-learn in c:\users\hp zbook\appdata\roaming\python\python311\site-packages (1.2.2)
Requirement already satisfied: numpy>=1.17.3 in c:\users\hp zbook\appdata\roaming\python\python311\site-packages (from sc
ikit-learn) (1.25.2)
Requirement already satisfied: scipy>=1.3.2 in c:\users\hp zbook\anaconda3\lib\site-packages (from scikit-learn) (1.11.1)
Requirement already satisfied: joblib>=1.1.1 in c:\users\hp zbook\anaconda3\lib\site-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\hp zbook\anaconda3\lib\site-packages (from scikit-learn)
(3.5.0)
Requirement already satisfied: seaborn in c:\users\hp zbook\appdata\roaming\python\python311\site-packages (0.13.1)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\hp zbook\appdata\roaming\python\python311\site-packages
(from seaborn) (1.25.2)
Requirement already satisfied: pandas>=1.2 in c:\users\hp zbook\appdata\roaming\python\python311\site-packages (from seab
orn) (1.5.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\hp zbook\appdata\roaming\python\python311\site-package
s (from seaborn) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\hp zbook\anaconda3\lib\site-packages (from matplotlib!=3.6.1,
>=3.4->seaborn) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\hp zbook\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.
4->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\hp zbook\anaconda3\lib\site-packages (from matplotlib!=3.6.
```

Note: *After running the above cell, kindly restart the notebook kernel and run all cells sequentially from the start again*.

In [3]:
```
# To help with reading and manipulation of data
import numpy as np
import pandas as pd

# Removes the limit for the number of displayed columns
pd.set_option("display.max_columns", None)
# Sets the limit for the number of displayed rows
pd.set_option("display.max_rows", 200)

# To help with data visualization
import matplotlib.pyplot as plt
import seaborn as sns
import scikitplot as skplt

# To split the data
from sklearn.model_selection import train_test_split

# To impute missing values
from sklearn.impute import SimpleImputer
from sklearn.impute import KNNImputer

# To preprocess the data for modelling
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import OneHotEncoder

# To build a logistic regression classifier
from sklearn.linear_model import LogisticRegression

# To build a Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree

# # To build different ensemble classifiers
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import StackingClassifier
from xgboost import XGBClassifier

# # To undersample and oversample the data
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler

# To tune a model
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV

# To create a pipeline for production
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.compose import ColumnTransformer
from sklearn.compose import make_column_selector as selector

# To get different performance metrics
import sklearn.metrics as metrics
from sklearn.metrics import (
    classification_report,
    confusion_matrix,
    recall_score,
    accuracy_score,
    precision_score,
    f1_score,
)

from scikitplot.metrics import (
    plot_confusion_matrix

)

# To suppress warnings
import warnings

warnings.filterwarnings("ignore")
```

Loading Data

In [4]:
```
churn_dataset = pd.read_csv("survey_response.csv")
```

In [5]:
```
# Checking the number of rows and columns in the data
churn_dataset.shape
```

Out[5]: (791, 14)

- The dataset has 791 rows and 14 columns

Data Overview

- Observations
- Sanity checks

In [6]:
```
# let's create a copy of the data
data = churn_dataset.copy()
```

In [7]:
```
# let's view the first 5 rows of the data
data.head()
```

Out[7]:

Timestamp	Gender	Work Sector	Educational Qualification	Age range	Do you have a permanent voters card?	Are you likely to vote when there is electoral violence around you?	Are you likely to prevent a "female" loved one from going to vote after violence occurs?	Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?	In your opinion, does violence impact the confidence of women in engaging in political activities?	Did you vote in 2023 General Elections?	If No, why not?	Have you ever witnessed any form of electoral violence during elections?	Have witne any fo harass on s me
04/04/2024 17:57	Female	Formal Sector (9-5 jobs, Professionals, Hybrid...	HND, B.Sc.	18-30	No	No	No	No	No	No	No PVC	No	
04/04/2024 18:05	Female	Formal Sector (9-5 jobs, Professionals, Hybrid...	HND, B.Sc.	18-30	No	No	Yes	Yes	Yes	No	Others	Yes	
04/04/2024 18:06	Female	Informal Sector (Artisans, Traders)	HND, B.Sc.	31-40	Yes	No	Yes	Yes	Yes	No	Unavailable (distance, health issues)	No	
04/04/2024 18:08	Male	Formal Sector (9-5 jobs, Professionals, Hybrid...	HND, B.Sc.	18-30	No	No	Yes	Yes	Yes	No	No PVC	Yes	
04/04/2024 18:10	Female	Informal Sector (Artisans, Traders)	HND, B.Sc.	18-30	Yes	No	Yes	Yes	Yes	Yes	Others	Yes	

In [83]:
```
# let's view the last 5 rows of the data
data.tail()
```

Out[83]:

	Timestamp	Gender	Work Sector	Educational Qualification	Age range	Do you have a permanent voters card?	Are you likely to vote when there is electoral violence around you?	Are you likely to prevent a "female" loved one from going to vote after violence occurs?	Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?	In your opinion, does violence impact the confidence of women in engaging in political activities?	Did you vote in 2023 General Elections?	If No, why not?	Have you eve witnesse any forn o electora violenc during elections
786	NaN	Female	Formal Sector (9-5 jobs, Professionals, Hybrid...	SSCE and below	31-40	Yes	No	Yes	Uncertain	Yes	Yes	Others	Uncertai
787	NaN	Female	Informal Sector	HND, B.Sc.	60 and	No	No	Yes	No	No	No	Work (Journalist, Health	Ye

In [84]:
```
# Timestamp consists of uniques ID for clients and hence will not add value to the modeling
data.drop(["Timestamp"], axis=1, inplace=True)
```

In [19]:
```
data.columns.tolist()
```

Out[19]:
```
['Gender',
 'Work Sector',
 'Educational Qualification',
 'Age range',
 'Do you have a permanent voters card?',
 'Are you likely to vote when there is electoral violence around you?',
 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?',
 'Do you believe that violence deters women from participating in political activities such as rallies, elections, and camp
aigns?',
 'In your opinion, does violence impact the confidence of women in engaging in political activities?',
 'Did you vote in 2023 General Elections?',
 'If No, why not?',
 'Have you ever witnessed any form of electoral violence during elections?',
 'Have you ever witnessed any form of harassment on social media?']
```

Data Type Conversions

In [27]:
```
# let's view the statistical summary of the numerical columns in the data
data.describe().T
```

Out[27]:

	count	unique	top	freq
Gender	791	2	Female	533
Work Sector	791	2	Formal Sector (9-5 jobs, Professionals, Hybrid...	492
Educational Qualification	791	7	HND, B.Sc.	438
Age range	791	6	18-30	380
Do you have a permanent voters card?	791	2	Yes	524
Are you likely to vote when there is electoral violence around you?	791	3	No	604
Are you likely to prevent a "female" loved one from going to vote after violence occurs?	791	3	Yes	691
Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?	791	3	Yes	603
In your opinion, does violence impact the confidence of women in engaging in political activities?	791	3	Yes	552
Did you vote in 2023 General Elections?	791	2	No	586
If No, why not?	791	5	Others	273
Have you ever witnessed any form of electoral violence during elections?	791	3	Yes	395
Have you ever witnessed any form of harassment on social media?	791	3	Yes	402

Observations:

- Dataset: the data has no missiing value
- Data type: all column has object has the datatype
- Colums: all the columns have categorical values

In [30]:

```
for i in data.describe(include=["object"]).columns:
    print("Unique values in", i, "are :")
    print(data[i].value_counts())
    print("*" * 50)
    print("*" * 50)
```

```
Unique values in Gender are :
Female    533
Male      258
Name: Gender, dtype: int64
**************************************************
**************************************************
Unique values in Work Sector are :
Formal Sector (9-5 jobs, Professionals, Hybrid jobs)    492
Informal Sector (Artisans, Traders)                     299
Name: Work Sector, dtype: int64
**************************************************
**************************************************
Unique values in Educational Qualification are :
HND, B.Sc.        438
Postgraduate      145
SSCE and below    140
ND, NCE            39
Mbbs in view       12
Btech               9
Undergraduate       8
Name: Educational Qualification, dtype: int64
**************************************************
**************************************************
Unique values in Age range are :
18-30           380
31-40           173
51-60           116
60 and above     71
41-50            50
51-61             1
Name: Age range, dtype: int64
**************************************************
**************************************************
Unique values in Do you have a permanent voters card? are :
Yes    524
No     267
Name: Do you have a permanent voters card?, dtype: int64
**************************************************
**************************************************
Unique values in Are you likely to vote when there is electoral violence around you? are :
No           604
Yes          140
Uncertain     47
Name: Are you likely to vote when there is electoral violence around you?, dtype: int64
**************************************************
**************************************************
Unique values in Are you likely to prevent a "female" loved one from going to vote after violence occurs? are :
Yes          691
No            91
Uncertain      9
Name: Are you likely to prevent a "female" loved one from going to vote after violence occurs?, dtype: int64
**************************************************
**************************************************
Unique values in Do you believe that violence deters women from participating in political activities such as rallies, elec
tions, and campaigns? are :
Yes          603
No           132
Uncertain     56
Name: Do you believe that violence deters women from participating in political activities such as rallies, elections, and
campaigns?, dtype: int64
**************************************************
**************************************************
Unique values in In your opinion, does violence impact the confidence of women in engaging in political activities? are :
Yes          552
No           193
Uncertain     46
Name: In your opinion, does violence impact the confidence of women in engaging in political activities?, dtype: int64
**************************************************
**************************************************
Unique values in Did you vote in 2023 General Elections? are :
No     586
Yes    205
Name: Did you vote in 2023 General Elections?, dtype: int64
**************************************************
**************************************************
Unique values in If No, why not? are :
Others                                                                  273
No PVC                                                                  218
Unavailable  (distance, health issues)                                  159
Electoral Violence                                                       83
Work (Journalist, Health officials, Security agents, Electoral officers)     58
Name: If No, why not?, dtype: int64
**************************************************
**************************************************
Unique values in Have you ever witnessed any form of electoral violence during elections? are :
Yes          395
No           343
Uncertain     53
Name: Have you ever witnessed any form of electoral violence during elections?, dtype: int64
**************************************************
**************************************************
Unique values in Have you ever witnessed any form of harassment on social media? are :
```

```
Yes          402
No           364
Uncertain     25
Name: Have you ever witnessed any form of harassment on social media?, dtype: int64
**************************************************
**************************************************
```

Observation

- there is no missing value

Exploratory Data Analysis (EDA)

- EDA is an important part of this project inorder to reveal hidden information from the data.
- It is important to investigate and understand the data better before building a model with it.

Some of the Questions Answered through the EDA:

1. How is the gender distributed?
2. What is the distribution of the Educational Qualification?
3. What is the distribution of the permanent voters card?
4. How does the change in likelyhood to vote when there is electoral violence vary by the gender
5. How does the witness of any form of electoral violence during elections vary by the gender
6. How does the witnessing any form of harassment on social media vary by the gender
7. What are the attributes that have a strong correlation with each other?

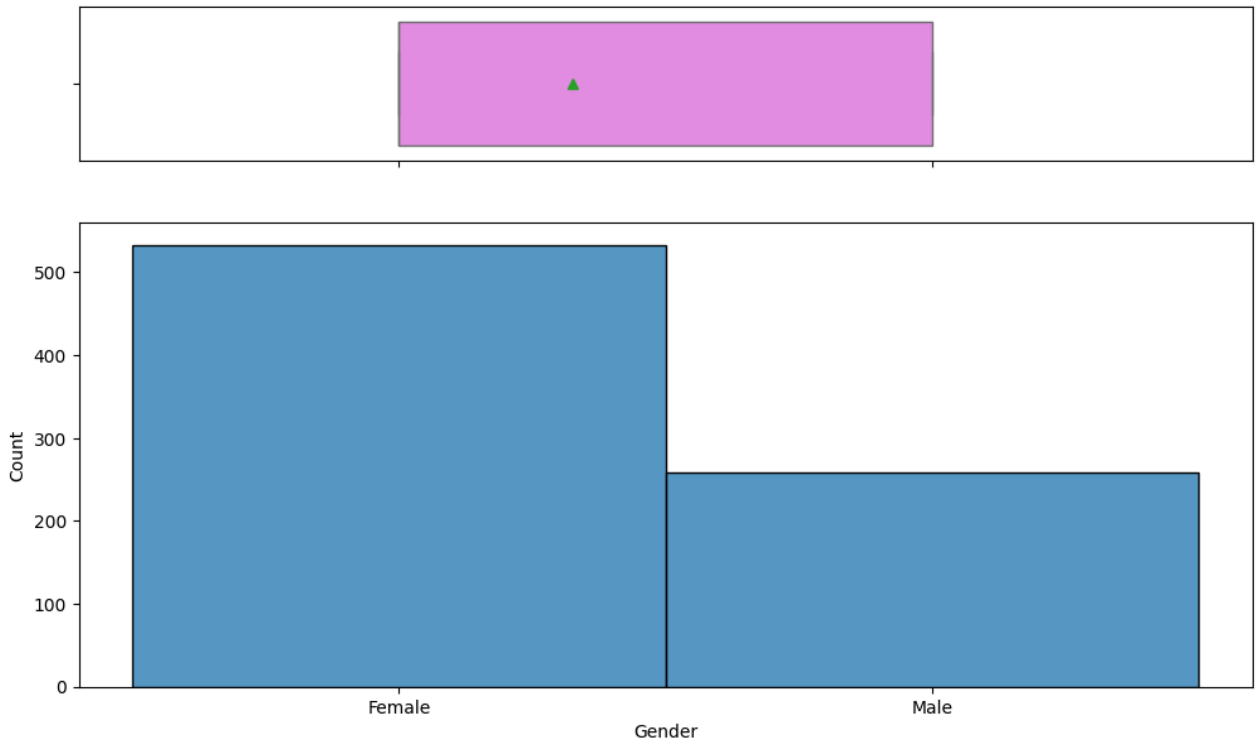
Univariate analysis

In [9]:

```
# function to plot a boxplot and a histogram along the same scale.


def histogram_boxplot(data, feature, figsize=(12, 7), kde=False, bins=None):
    """
    Boxplot and histogram combined

    data: dataframe
    feature: dataframe column
    figsize: size of figure (default (12,7))
    kde: whether to the show density curve (default False)
    bins: number of bins for histogram (default None)
    """
    f2, (ax_box2, ax_hist2) = plt.subplots(
        nrows=2,  # Number of rows of the subplot grid= 2
        sharex=True,  # x-axis will be shared among all subplots
        gridspec_kw={"height_ratios": (0.25, 0.75)},
        figsize=figsize,
    )  # creating the 2 subplots
    sns.boxplot(
        data=data, x=feature, ax=ax_box2, showmeans=True, color="violet"
    )  # boxplot will be created and a star will indicate the mean value of the column
    sns.histplot(
        data=data, x=feature, kde=kde, ax=ax_hist2, bins=bins, palette="winter"
    ) if bins else sns.histplot(
        data=data, x=feature, kde=kde, ax=ax_hist2
    )  # For histogram
#     ax_hist2.axvline(
#         data[feature].mean(), color="green", linestyle="--"
#     )  # Add mean to the histogram
#     ax_hist2.axvline(
#         data[feature].median(), color="black", linestyle="-"
#     )  # Add median to the histogram
```

Observations on Gender

In [32]:
```
1 histogram_boxplot(data, "Gender")
```

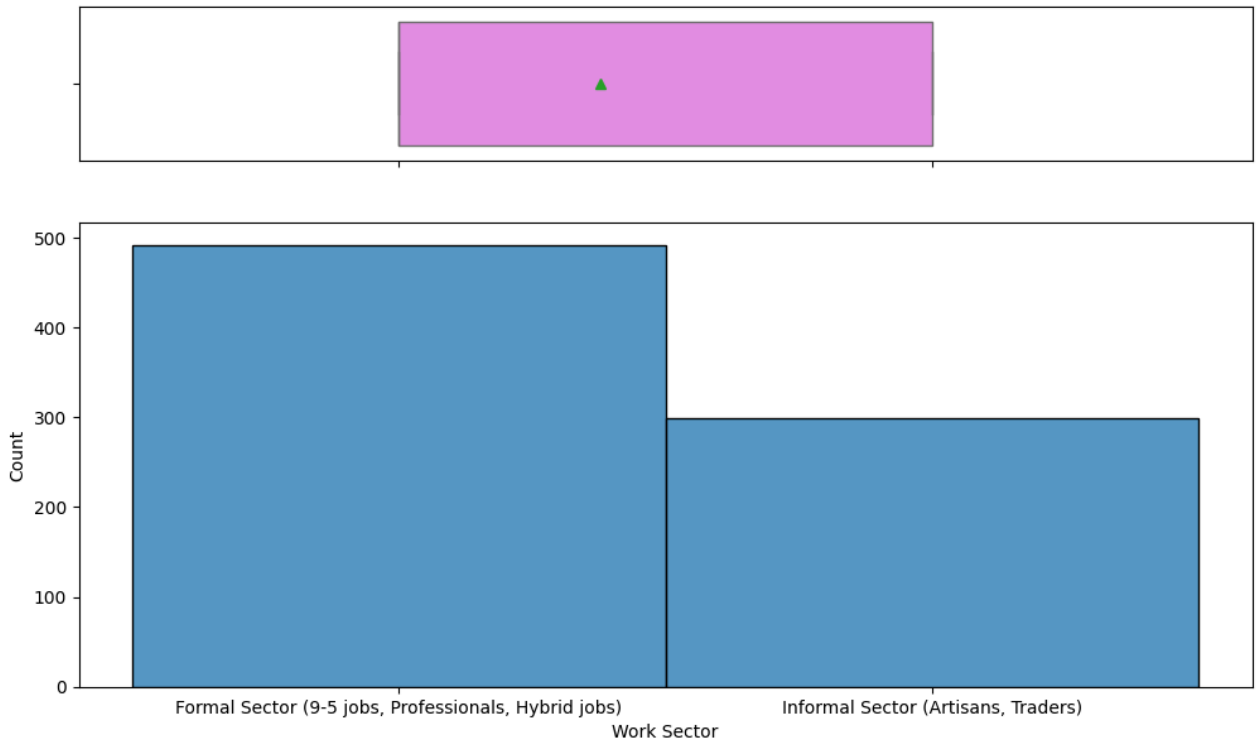


Observation

The are more female respondent than male

Observations on Work Sector

In [34]:
```
1 histogram_boxplot(data, "Work Sector")
```



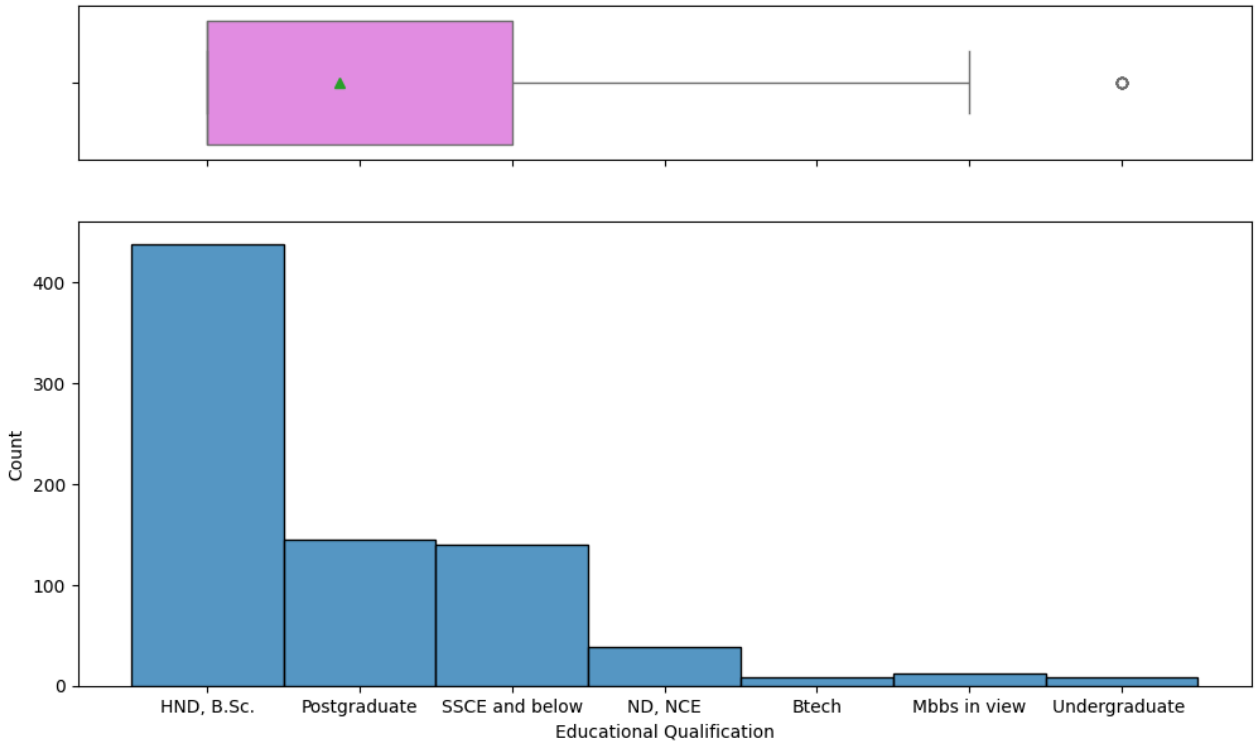
Observation

- Most of the respondent work in formal sector. That is the respondent is either works,
  1. on a 9:00AM to 5:00Pm schedule or
  2. as a professionals
  3. on jobs with hybrid mode

- Few Informal sectors (such as Artisans and Traders) responded.

Observations on Educational Qualification

In [35]:
```
1 histogram_boxplot(data, "Educational Qualification")
```

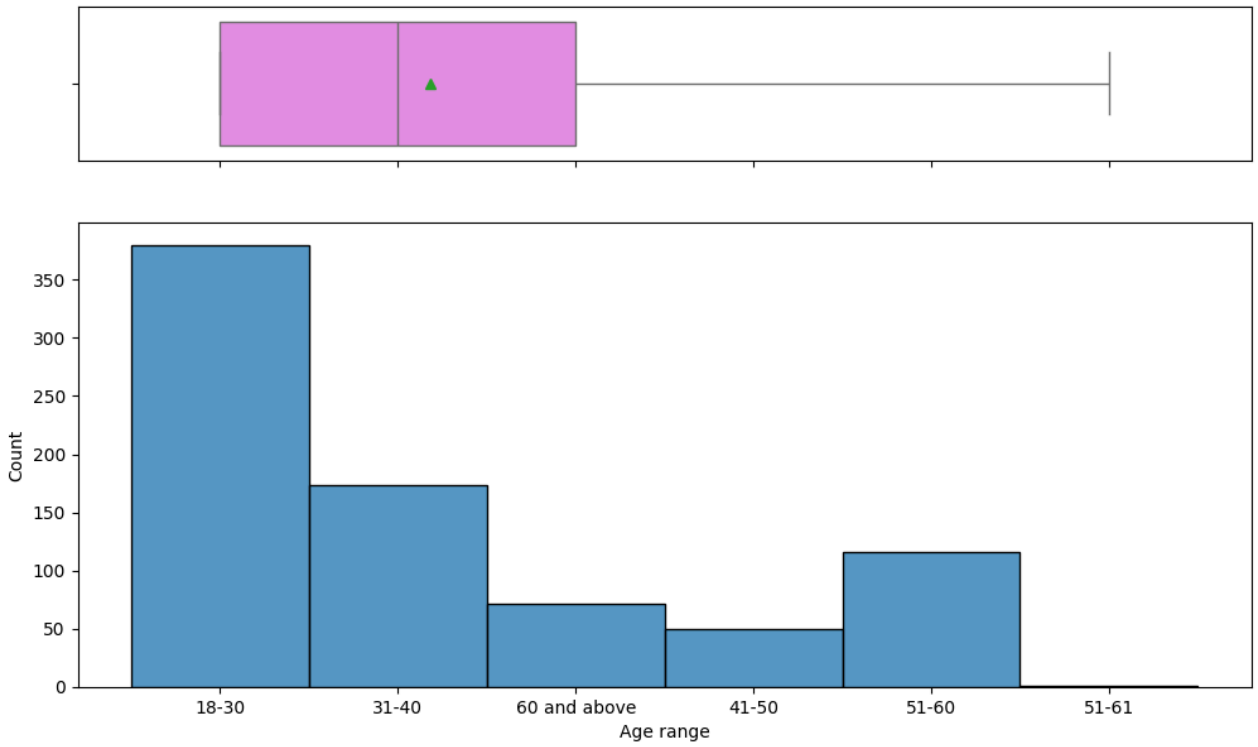


Observation

- More than 400 respondent holds HND or BSC Degree.
- Less than 200 respondent hold postgraduate degree.
- The respondent with ND or NCE or MBBS in View or BTech or Undergraduate are all less than 100.

Observations on Age Range

In [37]:
```
1 histogram_boxplot(data, "Age range")
```

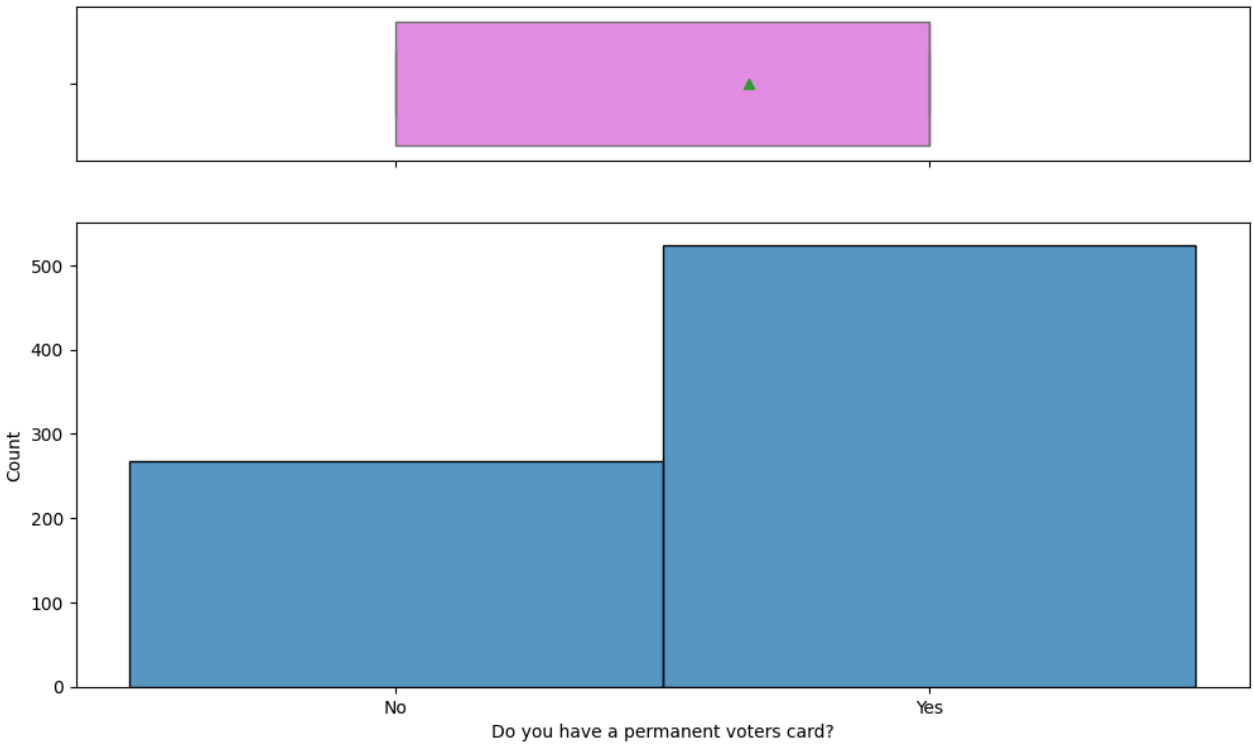


- 18-30 is the age range with the highest numebr of respondent follow by 31-40.

- 41-50 is the age range with the lowest number of respondent.

Observations on 'Respondent with permanent voters card?'

In [38]:
```
1 histogram_boxplot(data, "Do you have a permanent voters card?")
```

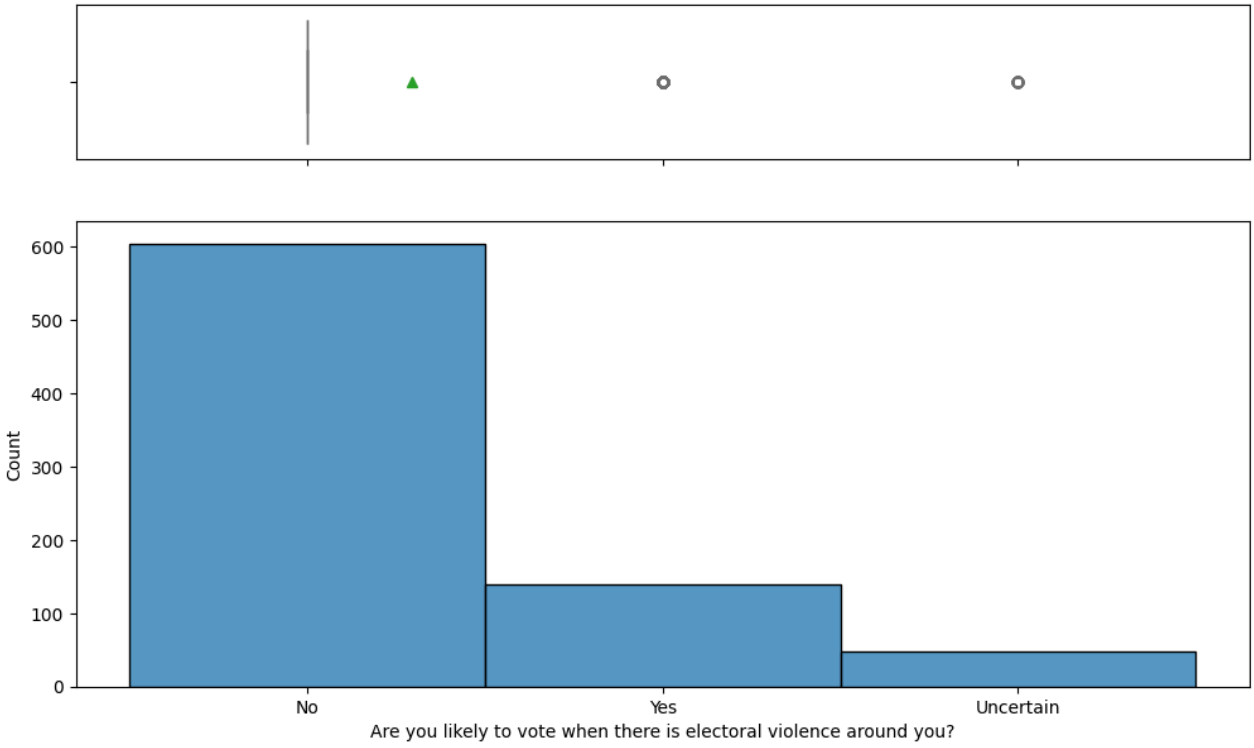


Observation

- About 70% of the respondent have permanent voters card

Observations on *Respondent that are likely to vote when there is electoral violence around you*?

In [39]:
```
1 histogram_boxplot(data, "Are you likely to vote when there is electoral violence around you?")
```

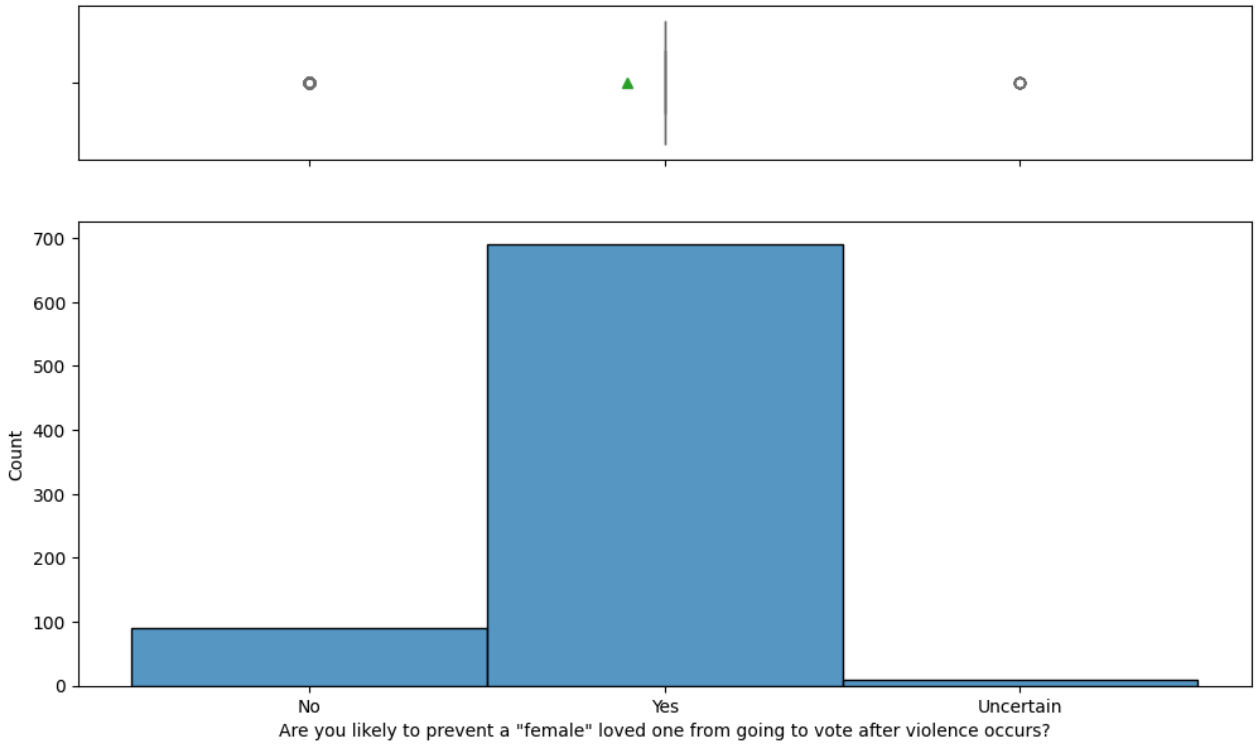


Observation

- around 80% of the respondent are not likely to vote when there is electoral violence around.

Observations on *Respondent that are likely to prevent a "female" loved one from going to vote after violence occurs*?

In [41]:
```
histogram_boxplot(data, 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?')
```

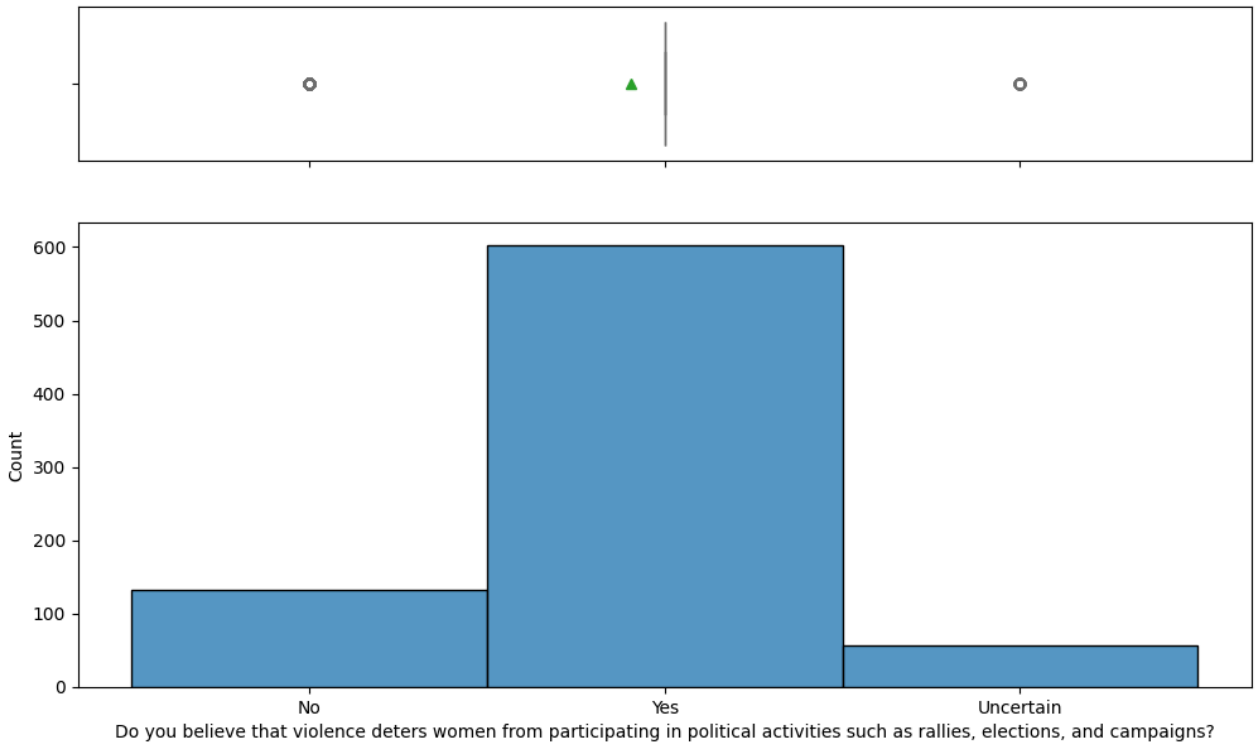


Observation

- 90% of the respondent indicate that they will prevent female loved ones from going to vote after violence occurs.

Observations on Respondent that believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?

In [42]:
```
1 histogram_boxplot(data, "Do you believe that violence deters women from participating in political activities such as ra
```

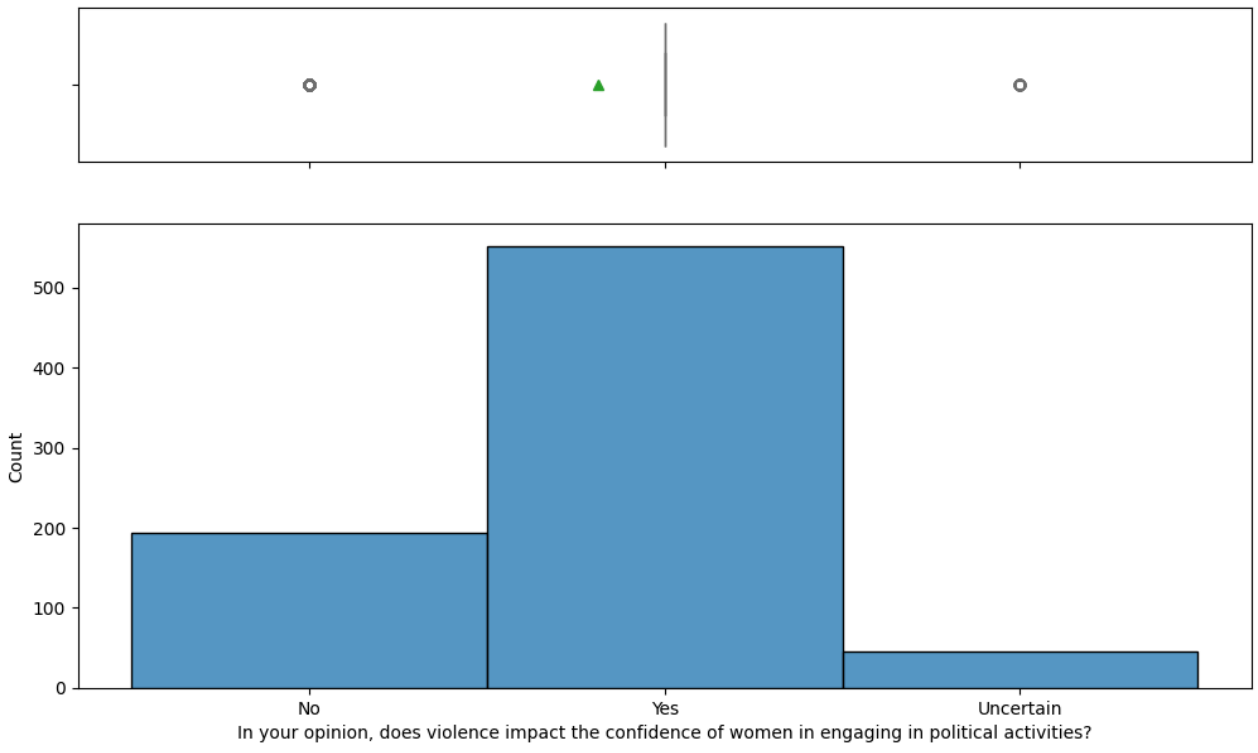


Observation

- Majority of the respondent believe that violenve deters women/female from participating in polical activities such as rallies, elections and campaigns

Observations on Respondent who have the opinoin that violence impact the confidence of women in engaging in political activities?

In [43]:
```
1 histogram_boxplot(data, "In your opinion, does violence impact the confidence of women in engaging in political activiti
```

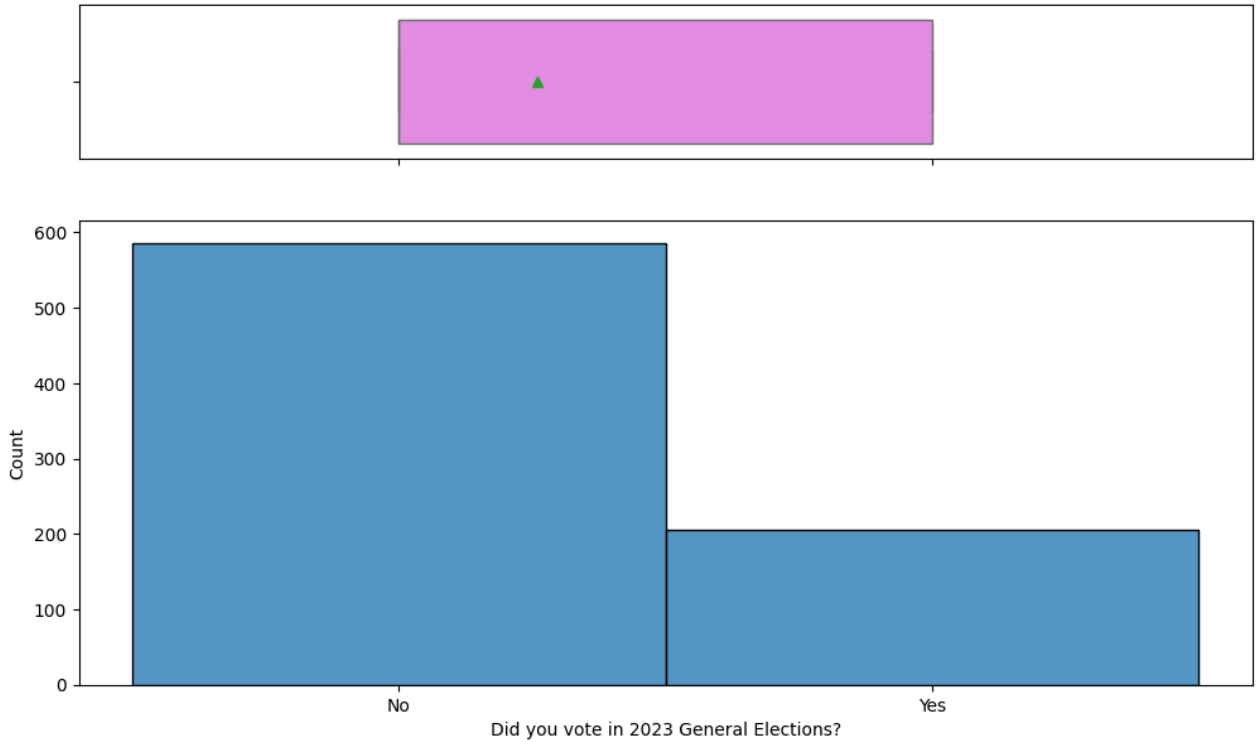


Observation

- Majority of the respondent beleive violence impact the confidence of women in engaging in political activities

Observations on Respondent that vote in 2023 General Elections?

In [44]:
```
1 histogram_boxplot(data, "Did you vote in 2023 General Elections?")
```

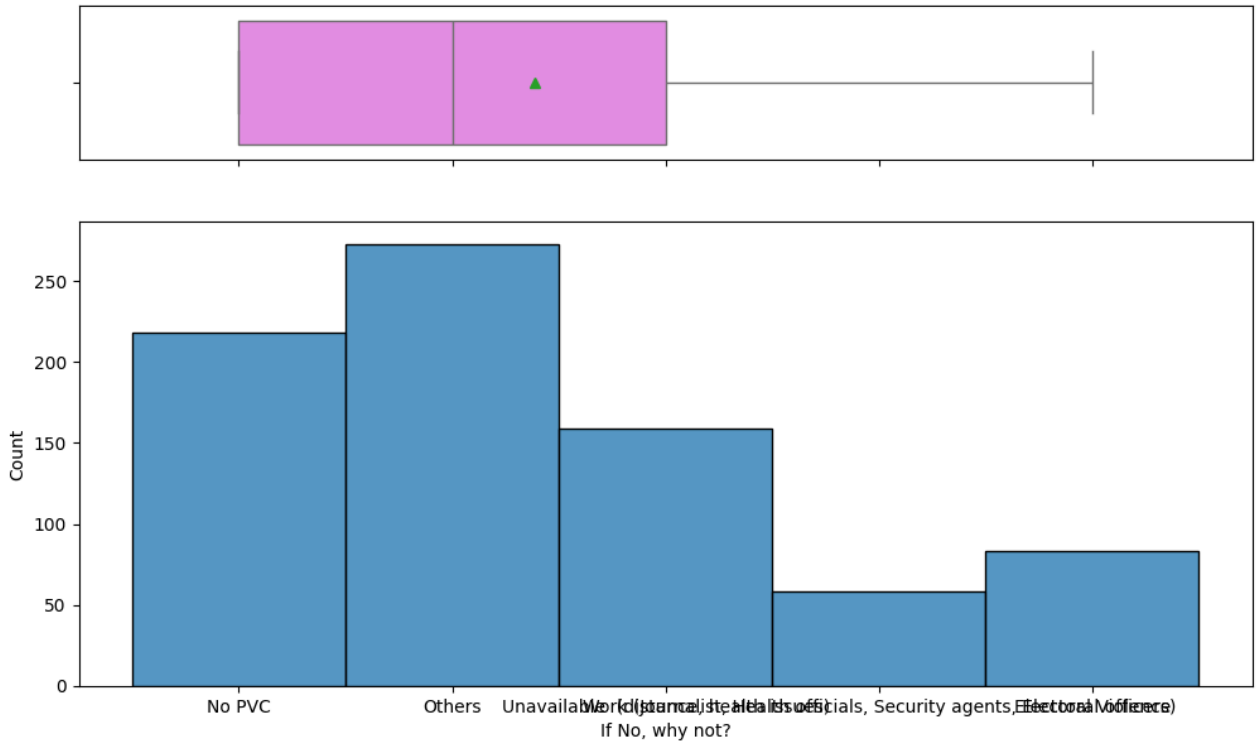


Observation

- Less that 300 respondent did not vote in 2023 General Election

Observation on Why Some Respondent did Not Vote in 2023 general election

In [52]:
```
1 histogram_boxplot(data, "If No, why not?")
```



Observation

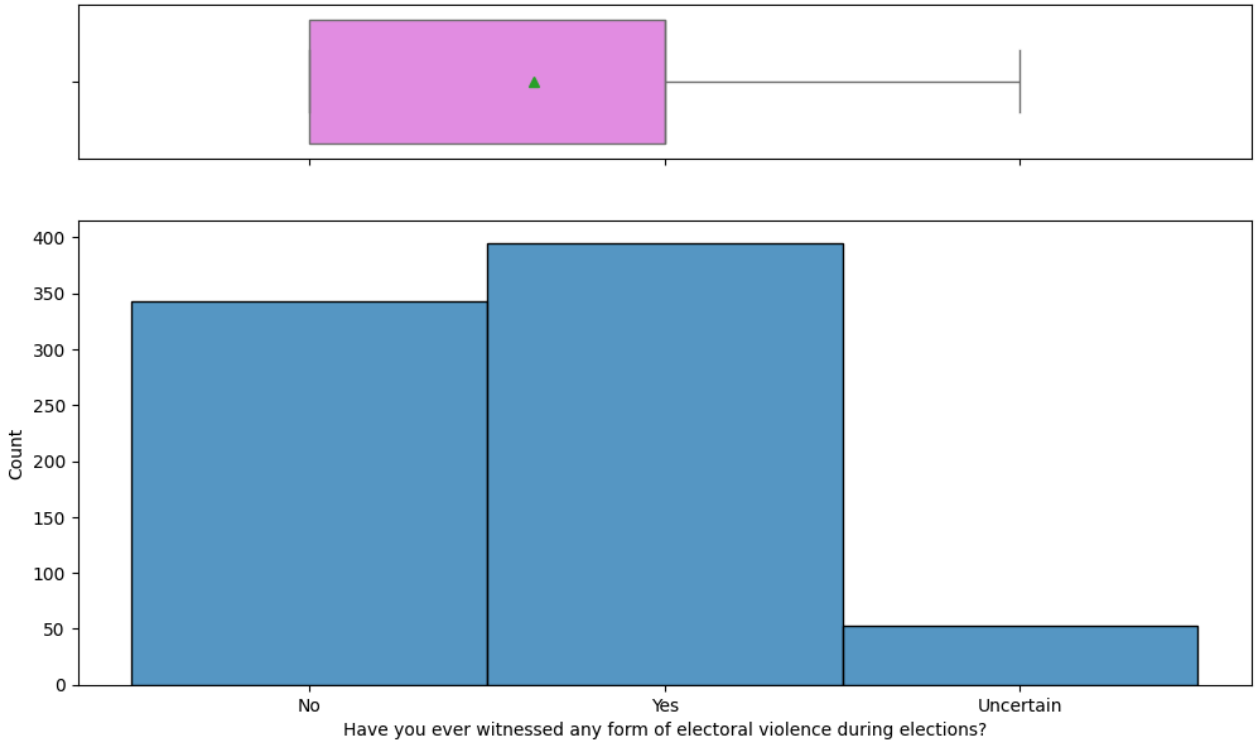
- Majority of the respondent that did not vote are influenced by

1. Other reason best known to them
2. Lack of Permanent Voters Card (PVC) and
3. Unavailability (such as distance, health issues)

- Less than 100 respondent did not vote due to Electoral Violence
- Work (Journalist, Health officials, Security agents, Electoral officers) is reason for the least number of respondent that did not vote

Observation on Respondent who has ever witnessed any form of electoral violence during elections?

In [54]:
```
histogram_boxplot(data, "Have you ever witnessed any form of electoral violence during elections?")
```

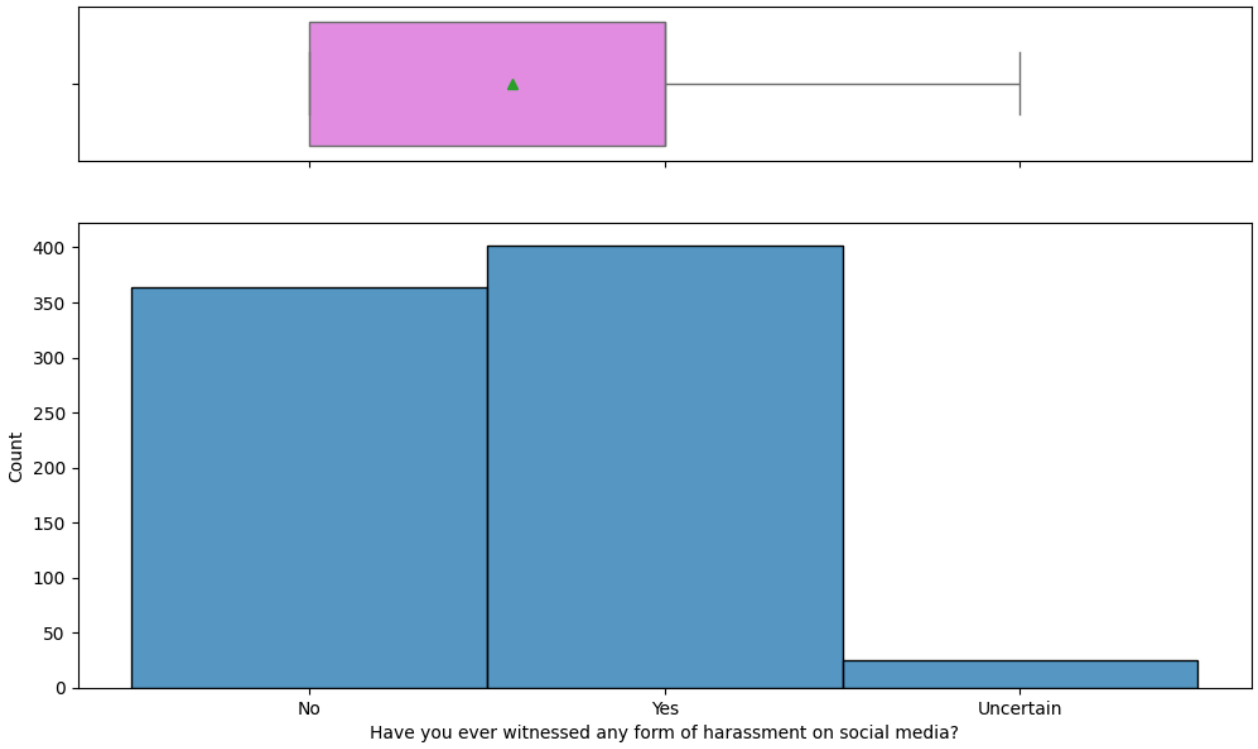


Observation

- Mjority od the respondent has witness at least a particular rform of electoral violence

Observation on Respondent who has ever witnessed any form of harassment on social media?

In [10]:
```
histogram_boxplot(data, "Have you ever witnessed any form of harassment on social media?")
```



Observation

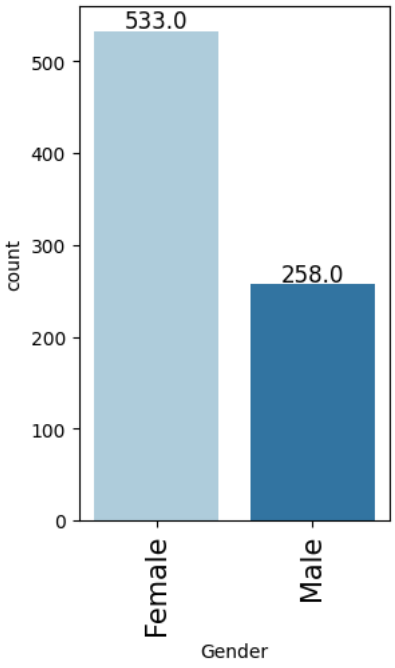
- Majority of the resondent has witnessed at least a particular form of harrasment on socia media

In [12]:

```
# function to create labeled barplots


def labeled_barplot(data, feature, perc=False, n=None):
    """
    Barplot with percentage at the top

    data: dataframe
    feature: dataframe column
    perc: whether to display percentages instead of count (default is False)
    n: displays the top n category levels (default is None, i.e., display all levels)
    """

    total = len(data[feature])  # Length of the column
    count = data[feature].nunique()
    if n is None:
        plt.figure(figsize=(count + 1, 5))
    else:
        plt.figure(figsize=(n + 1, 5))

    plt.xticks(rotation=90, fontsize=15)
    ax = sns.countplot(
        data=data,
        x=feature,
        palette="Paired",
        order=data[feature].value_counts().index[:n].sort_values(),
    )

    for p in ax.patches:
        if perc == True:
            label = "{:.1f}%".format(
                100 * p.get_height() / total
            )  # percentage of each class of the category
        else:
            label = p.get_height()  # count of each level of the category

        x = p.get_x() + p.get_width() / 2  # width of the plot
        y = p.get_height()  # height of the plot

        ax.annotate(
            label,
            (x, y),
            ha="center",
            va="center",
            size=12,
            xytext=(0, 5),
            textcoords="offset points",
        )  # annotate the percentage

    plt.show()  # show the plot
```

Observations on Gender

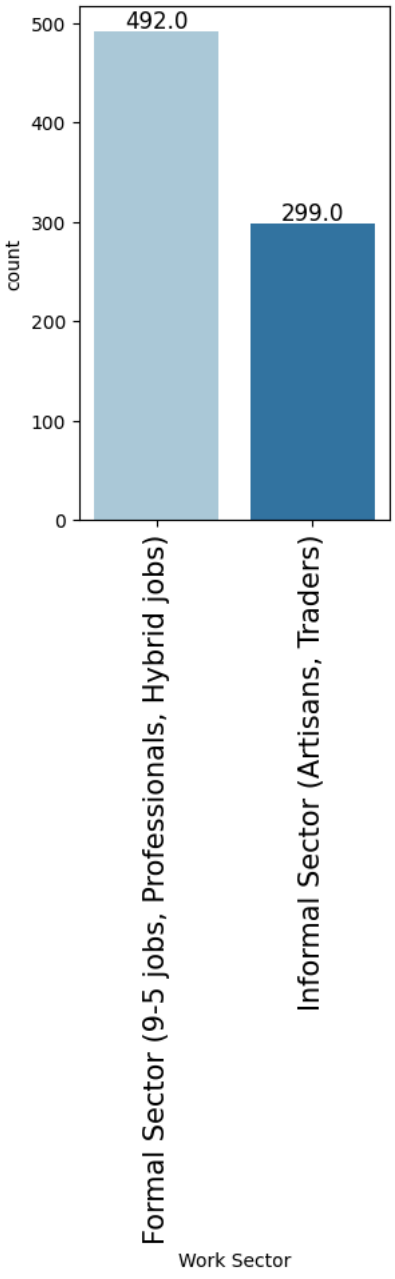
In [57]:
```
1 labeled_barplot(data, "Gender")
```



- The female respondent is more than the male respondent.

Observations on Work Sector

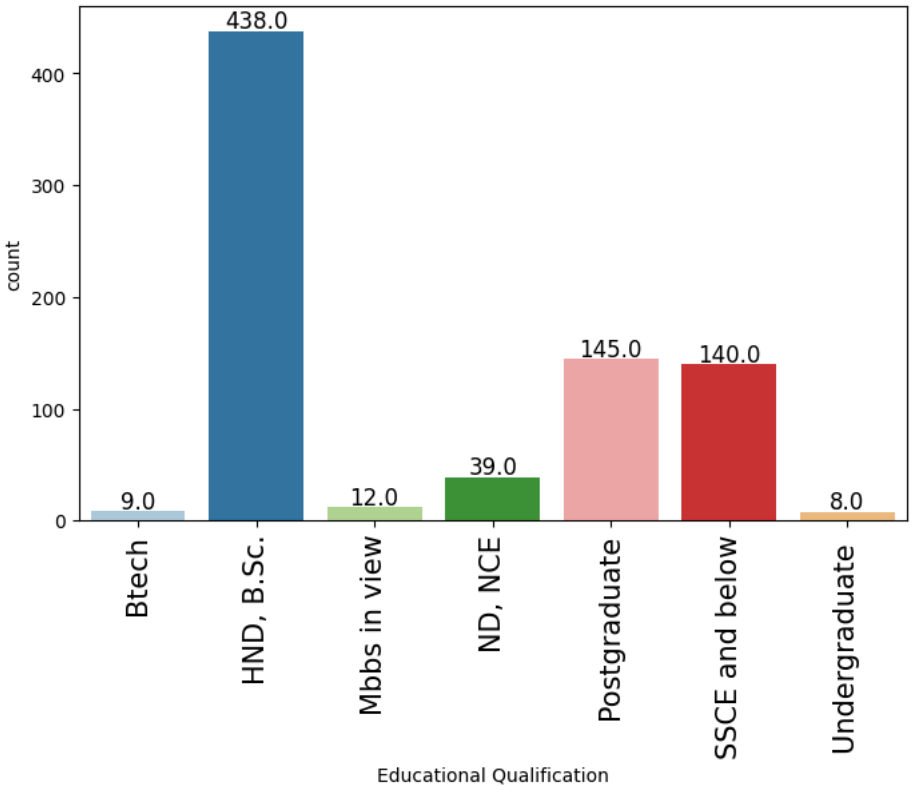
In [58]:
```
labeled_barplot(data, "Work Sector")
```



Type *Markdown* and LaTeX: α^2

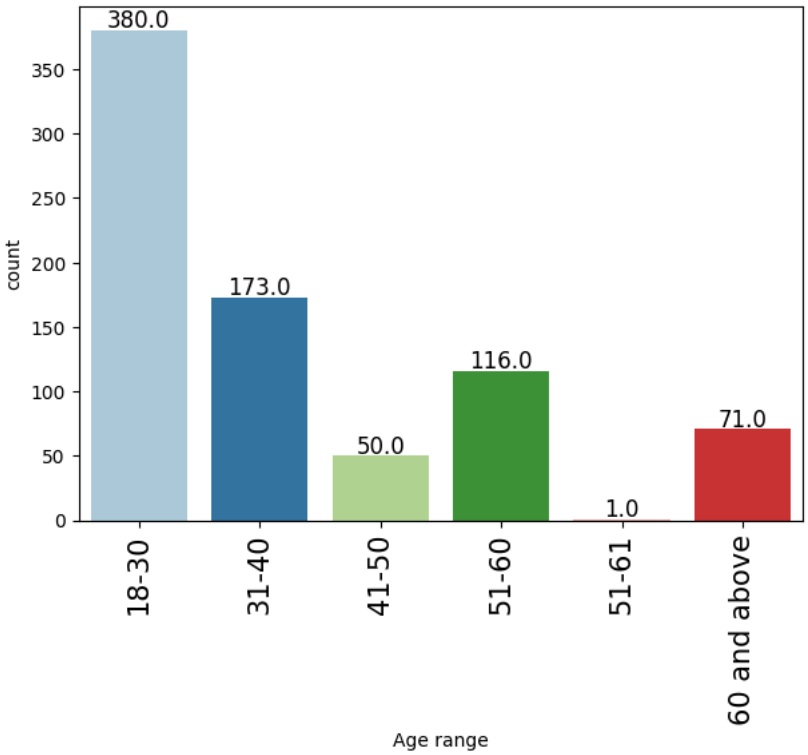
Observations on Educational Qualification

In [59]:
```
labeled_barplot(data, "Educational Qualification")
```



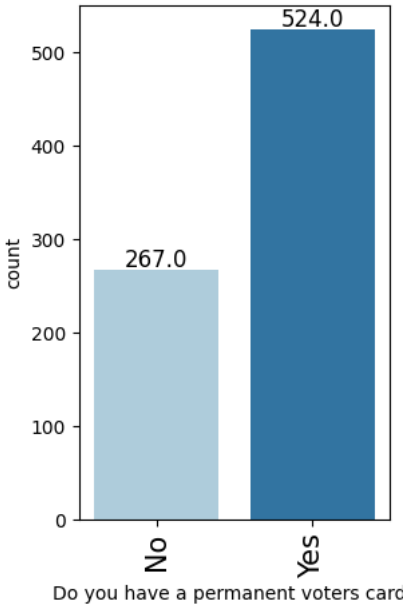
Observations on Age range

In [60]:
```
labeled_barplot(data, "Age range")
```



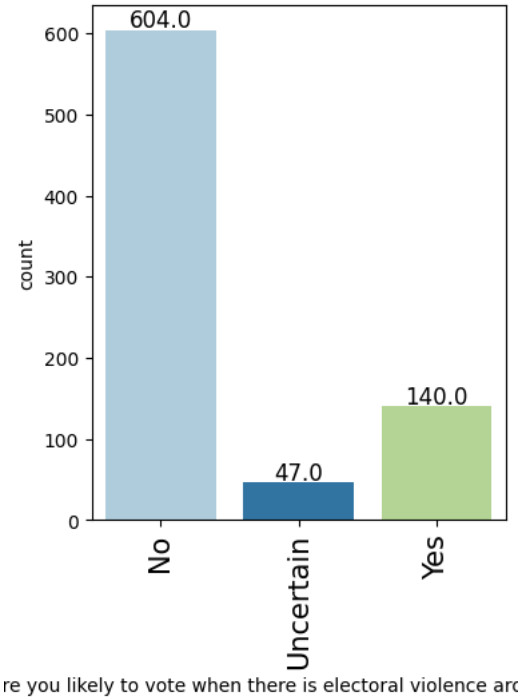
Observations on Do you have a permanent voters card?

In [61]:
```
labeled_barplot(data, "Do you have a permanent voters card?")
```



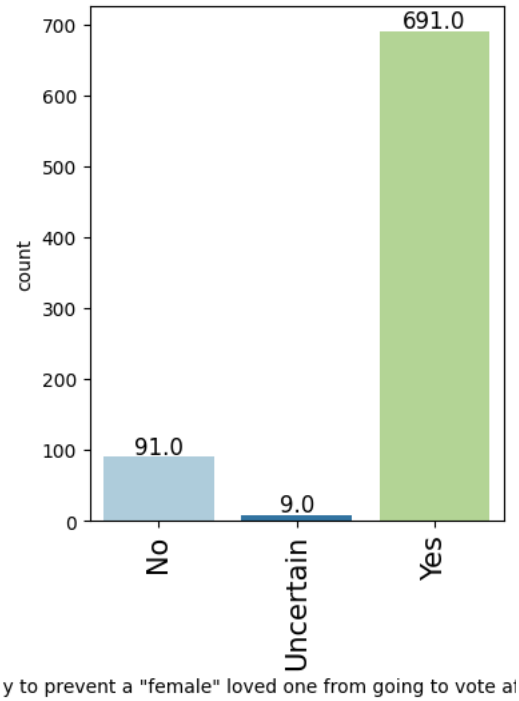
Observations on Are you likely to vote when there is electoral violence around you?

In [62]:
```
labeled_barplot(data, "Are you likely to vote when there is electoral violence around you?")
```



Observations on Are you likely to prevent a "female" loved one from going to vote after violence occurs?

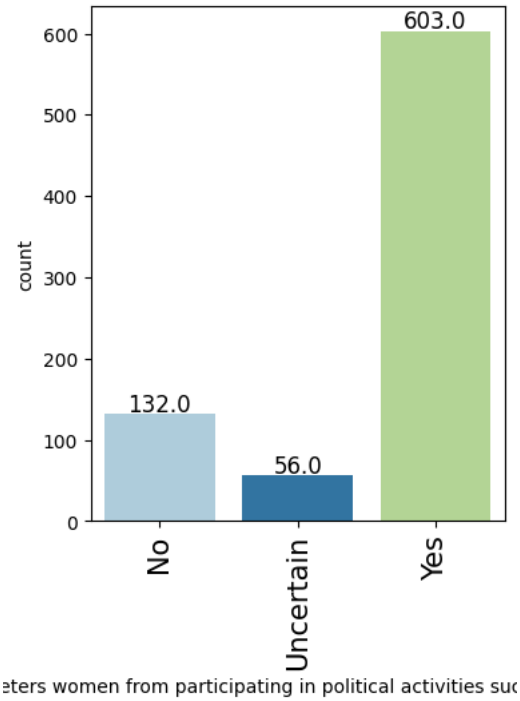
In [63]:
```
labeled_barplot(data, 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?')
```



y to prevent a "female" loved one from going to vote af

Observations on Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?

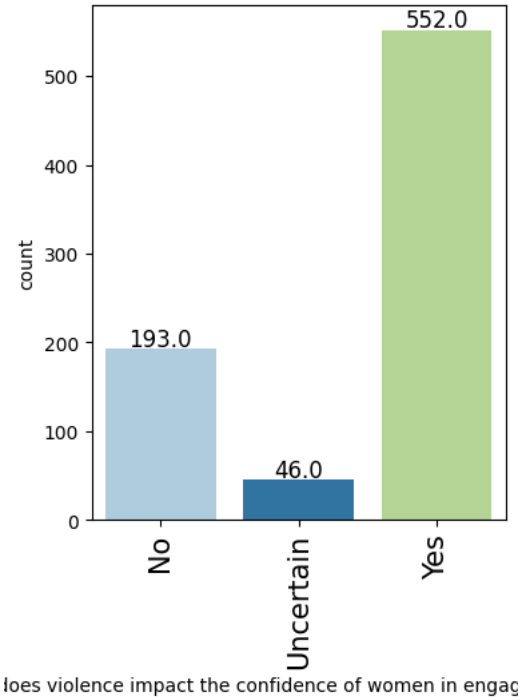
In [64]:
```
labeled_barplot(data, "Do you believe that violence deters women from participating in political activities such as rall
```



eters women from participating in political activities suc

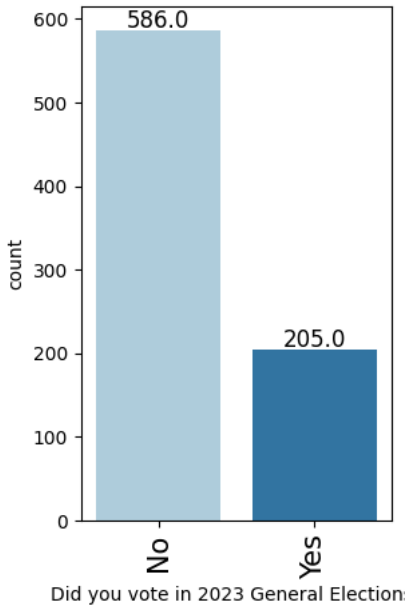
Observations on In your opinion, does violence impact the confidence of women in engaging in political activities?

In [65]:
```
labeled_barplot(data, "In your opinion, does violence impact the confidence of women in engaging in political activities
```



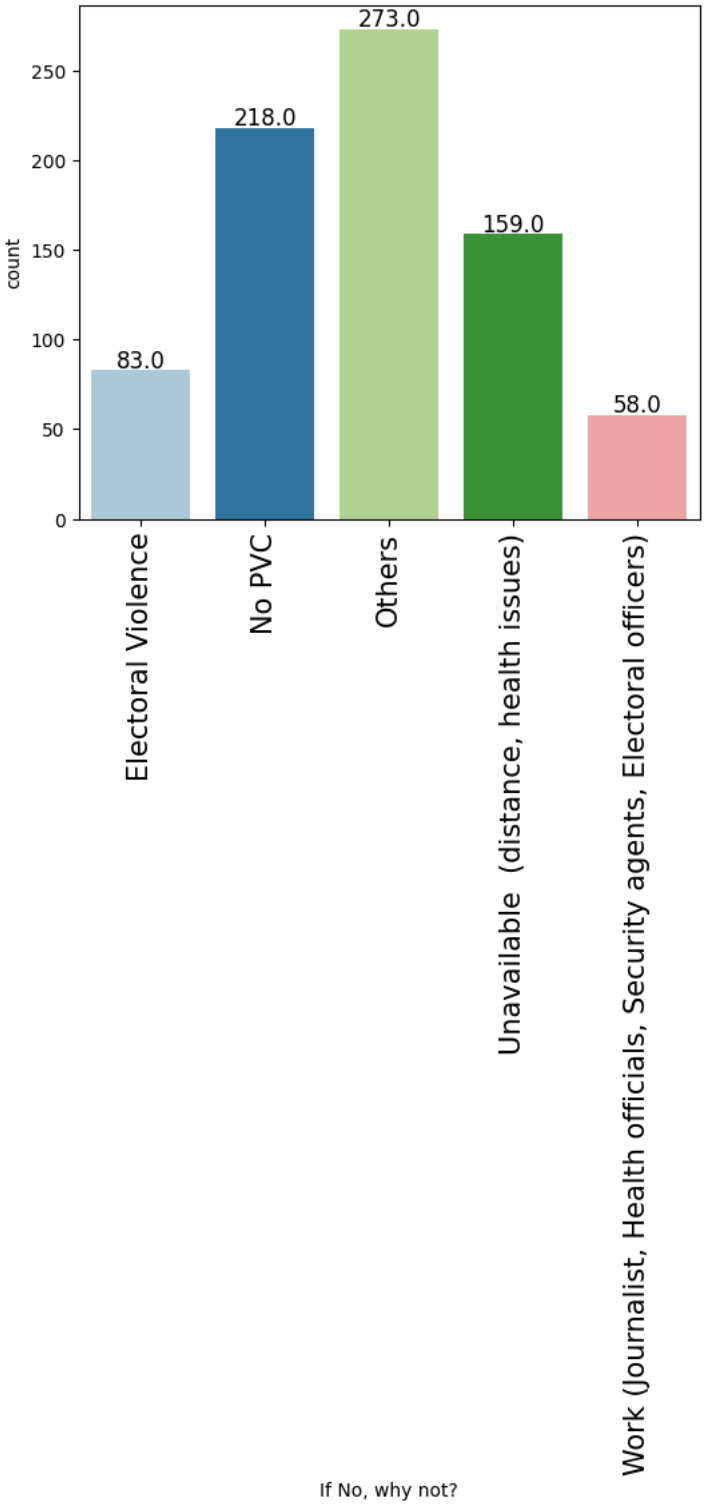
Observations on Did you vote in 2023 General Elections?

In [66]:
```
labeled_barplot(data, "Did you vote in 2023 General Elections?")
```



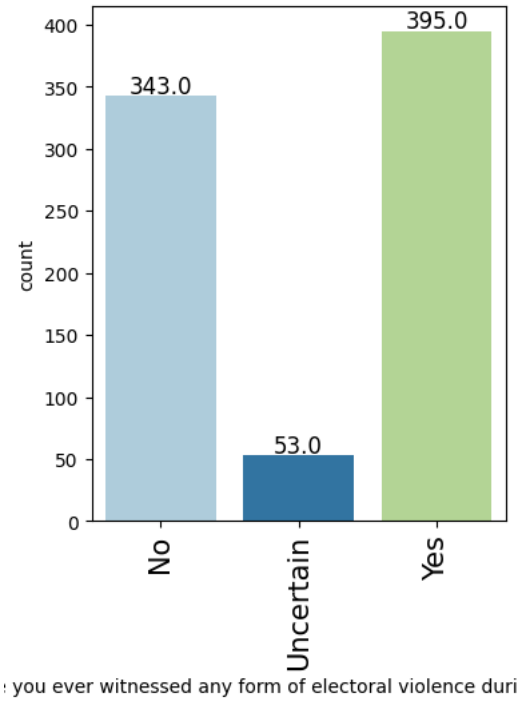
Observations on If No, why not?

In [67]:
```
1 labeled_barplot(data, "If No, why not?")
```



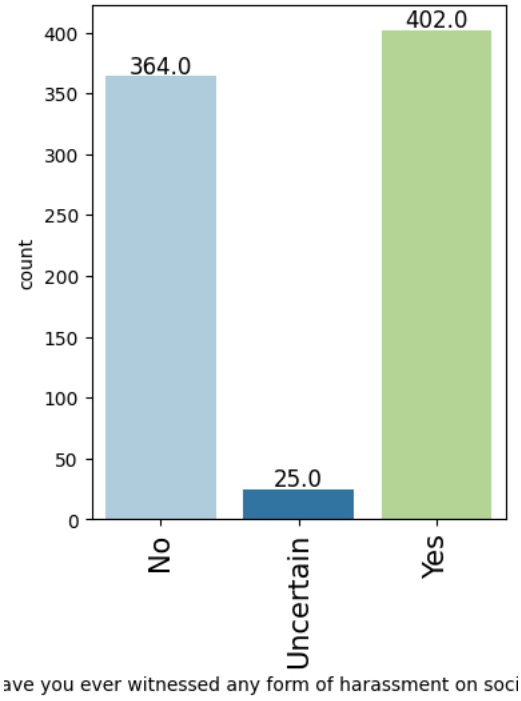
Observations on Have you ever witnessed any form of electoral violence during elections?

In [68]:
```
labeled_barplot(data, "Have you ever witnessed any form of electoral violence during elections?")
```



Observations on Have you ever witnessed any form of harassment on social media?

In [69]:
```
labeled_barplot(data, "Have you ever witnessed any form of harassment on social media?")
```



Bivariate Analysis

In [8]:

```
for i in data.describe(include=["object"]).columns:
    print("Unique values in", i, "are :")
    print(data[i].value_counts())
    print("*" * 50)
    print("*" * 50)
```

```
Unique values in Gender are :
Female    533
Male      258
Name: Gender, dtype: int64
**************************************************
**************************************************
Unique values in Work Sector are :
Formal Sector (9-5 jobs, Professionals, Hybrid jobs)    492
Informal Sector (Artisans, Traders)                     299
Name: Work Sector, dtype: int64
**************************************************
**************************************************
Unique values in Educational Qualification are :
HND, B.Sc.        438
Postgraduate      145
SSCE and below    140
ND, NCE            39
Mbbs in view       12
Btech               9
Undergraduate       8
Name: Educational Qualification, dtype: int64
**************************************************
**************************************************
Unique values in Age range are :
18-30           380
31-40           173
51-60           116
60 and above     71
41-50            50
51-61             1
Name: Age range, dtype: int64
**************************************************
**************************************************
Unique values in Do you have a permanent voters card? are :
Yes    524
No     267
Name: Do you have a permanent voters card?, dtype: int64
**************************************************
**************************************************
Unique values in Are you likely to vote when there is electoral violence around you? are :
No           604
Yes          140
Uncertain     47
Name: Are you likely to vote when there is electoral violence around you?, dtype: int64
**************************************************
**************************************************
Unique values in Are you likely to prevent a "female" loved one from going to vote after violence occurs? are :
Yes          691
No            91
Uncertain      9
Name: Are you likely to prevent a "female" loved one from going to vote after violence occurs?, dtype: int64
**************************************************
**************************************************
Unique values in Do you believe that violence deters women from participating in political activities such as rallies, elec
tions, and campaigns? are :
Yes          603
No           132
Uncertain     56
Name: Do you believe that violence deters women from participating in political activities such as rallies, elections, and
campaigns?, dtype: int64
**************************************************
**************************************************
Unique values in In your opinion, does violence impact the confidence of women in engaging in political activities? are :
Yes          552
No           193
Uncertain     46
Name: In your opinion, does violence impact the confidence of women in engaging in political activities?, dtype: int64
**************************************************
**************************************************
Unique values in Did you vote in 2023 General Elections? are :
No     586
Yes    205
Name: Did you vote in 2023 General Elections?, dtype: int64
**************************************************
**************************************************
Unique values in If No, why not? are :
Others                                                                  273
No PVC                                                                  218
Unavailable  (distance, health issues)                                  159
Electoral Violence                                                       83
Work (Journalist, Health officials, Security agents, Electoral officers)     58
Name: If No, why not?, dtype: int64
**************************************************
**************************************************
Unique values in Have you ever witnessed any form of electoral violence during elections? are :
Yes          395
No           343
Uncertain     53
Name: Have you ever witnessed any form of electoral violence during elections?, dtype: int64
**************************************************
**************************************************
Unique values in Have you ever witnessed any form of harassment on social media? are :
```

```
Yes          402
No           364
Uncertain     25
Name: Have you ever witnessed any form of harassment on social media?, dtype: int64
**************************************************
**************************************************
```

In [27]:
```
1 data.columns
```

Out[27]:
```
Index(['Gender', 'Work Sector', 'Educational Qualification', 'Age range',
       'Do you have a permanent voters card?',
       'Are you likely to vote when there is electoral violence around you?',
       'Are you likely to prevent a "female" loved one from going to vote after violence occurs?',
       'Do you believe that violence deters women from participating in political activities such as rallies, elections, an
d campaigns?',
       'In your opinion, does violence impact the confidence of women in engaging in political activities?',
       'Did you vote in 2023 General Elections?', 'If No, why not?',
       'Have you ever witnessed any form of electoral violence during elections?',
       'Have you ever witnessed any form of harassment on social media?'],
      dtype='object')
```

In [67]:

```
## Encoding Existing and Attrited customers to 0 and 1 respectively, for analysis.
# data["Attrition_F"].replace("Existing Customer", 0, inplace=True)
# data["Attrition_Flag"].replace("Attrited Customer", 1, inplace=True)

data["Gender"].replace("Male", 1, inplace=True)
data["Gender"].replace("Female", 0, inplace=True)

data["Work Sector"].replace("Formal Sector (9-5 jobs, Professionals, Hybrid jobs)", 1, inplace=True)
data["Work Sector"].replace("Informal Sector (Artisans, Traders)", 0, inplace=True)

data["Do you have a permanent voters card?"].replace("Yes", 1, inplace=True)
data["Do you have a permanent voters card?"].replace("No", 0, inplace=True)

data["Did you vote in 2023 General Elections?"].replace("Yes", 1, inplace=True)
data["Did you vote in 2023 General Elections?"].replace("No", 0, inplace=True)

data["Educational Qualification"].replace("Postgraduate", 20, inplace=True)
data["Educational Qualification"].replace("Mbbs in view", 15.5, inplace=True)
data["Educational Qualification"].replace("Btech", 10.5, inplace=True)
data["Educational Qualification"].replace("HND, B.Sc.", 10, inplace=True)
data["Educational Qualification"].replace("ND, NCE", 9, inplace=True)
data["Educational Qualification"].replace("Undergraduate ", "Undergraduate", inplace=True)
data["Educational Qualification"].replace("Undergraduate", 8, inplace=True)
data["Educational Qualification"].replace("SSCE and below", 5, inplace=True)
data["Educational Qualification"].astype('float64')

data["Age range"].replace("18-30", 24, inplace=True)
data["Age range"].replace("31-40", 35.5, inplace=True)
data["Age range"].replace("51-60", 55.5, inplace=True)
data["Age range"].replace("41-50", 45.5, inplace=True)
data["Age range"].replace("51-61", 56, inplace=True)
data["Age range"].replace("60 and above", 80, inplace=True)
data["Age range"].astype('float64')

data["Are you likely to vote when there is electoral violence around you?"].replace("No", 0, inplace=True)
data["Are you likely to vote when there is electoral violence around you?"].replace("Yes", 1, inplace=True)
data["Are you likely to vote when there is electoral violence around you?"].replace("Uncertain", 0.5, inplace=True)
data["Are you likely to vote when there is electoral violence around you?"].astype('float64')

data['Are you likely to prevent a "female" loved one from going to vote after violence occurs?'].replace("No", 0, inplac
data['Are you likely to prevent a "female" loved one from going to vote after violence occurs?'].replace("Yes", 1, inpla
data['Are you likely to prevent a "female" loved one from going to vote after violence occurs?'].replace("Uncertain", 0.
data['Are you likely to prevent a "female" loved one from going to vote after violence occurs?'].astype('float64')

data["Do you believe that violence deters women from participating in political activities such as rallies, elections, a
data["Do you believe that violence deters women from participating in political activities such as rallies, elections, a
data["Do you believe that violence deters women from participating in political activities such as rallies, elections, a
data["Do you believe that violence deters women from participating in political activities such as rallies, elections, a

data["If No, why not?"].replace("Others", 0.3, inplace=True)
data["If No, why not?"].replace("No PVC", 0, inplace=True)
data["If No, why not?"].replace("Unavailable  (distance, health issues)", 0.5, inplace=True)
data["If No, why not?"].replace("Electoral Violence", 1, inplace=True)
data["If No, why not?"].replace("Work (Journalist, Health officials, Security agents, Electoral officers)", 0.5, inplace
data["If No, why not?"].astype('float64')

data["Have you ever witnessed any form of electoral violence during elections?"].replace("Yes", 1, inplace=True)
data["Have you ever witnessed any form of electoral violence during elections?"].replace("No", 0, inplace=True)
data["Have you ever witnessed any form of electoral violence during elections?"].replace("Uncertain", 0.5, inplace=True)
data["Have you ever witnessed any form of electoral violence during elections?"].astype('float64')

data["Have you ever witnessed any form of harassment on social media?"].replace("Yes", 1, inplace=True)
data["Have you ever witnessed any form of harassment on social media?"].replace("No", 0, inplace=True)
data["Have you ever witnessed any form of harassment on social media?"].replace("Uncertain", 0.5, inplace=True)
data["Have you ever witnessed any form of harassment on social media?"].astype('float64')
```

Out[67]:

```
0      0.0
1      1.0
2      0.0
3      0.0
4      1.0
      ...
786    1.0
787    0.0
788    0.0
789    0.0
790    0.0
Name: Have you ever witnessed any form of harassment on social media?, Length: 791, dtype: float64
```

In [73]:
```
dataCorr = data.copy()

dataCorr.rename(columns = {'Educational Qualification':'Edu. Qlf.'}, inplace = True)
dataCorr.rename(columns = {'Do you have a permanent voters card?':'PVC'}, inplace = True)
dataCorr.rename(columns = {'Did you vote in 2023 General Elections?':'Vote in 2023 Gen. Elec.'}, inplace = True)
dataCorr.rename(columns = {'Are you likely to vote when there is electoral violence around you?':'Vote During Elec. Vio.
dataCorr.rename(columns = {'Are you likely to prevent a "female" loved one from going to vote after violence occurs?':'A
dataCorr.rename(columns = {'Do you believe that violence deters women from participating in political activities such as
dataCorr.rename(columns = {'In your opinion, does violence impact the confidence of women in engaging in political activ
dataCorr.rename(columns = {'Have you ever witnessed any form of electoral violence during elections?':'Witnessed any Ele
dataCorr.rename(columns = {'Have you ever witnessed any form of harassment on social media?':'Witnessed Haras. Social Me

dataCorr.columns
```

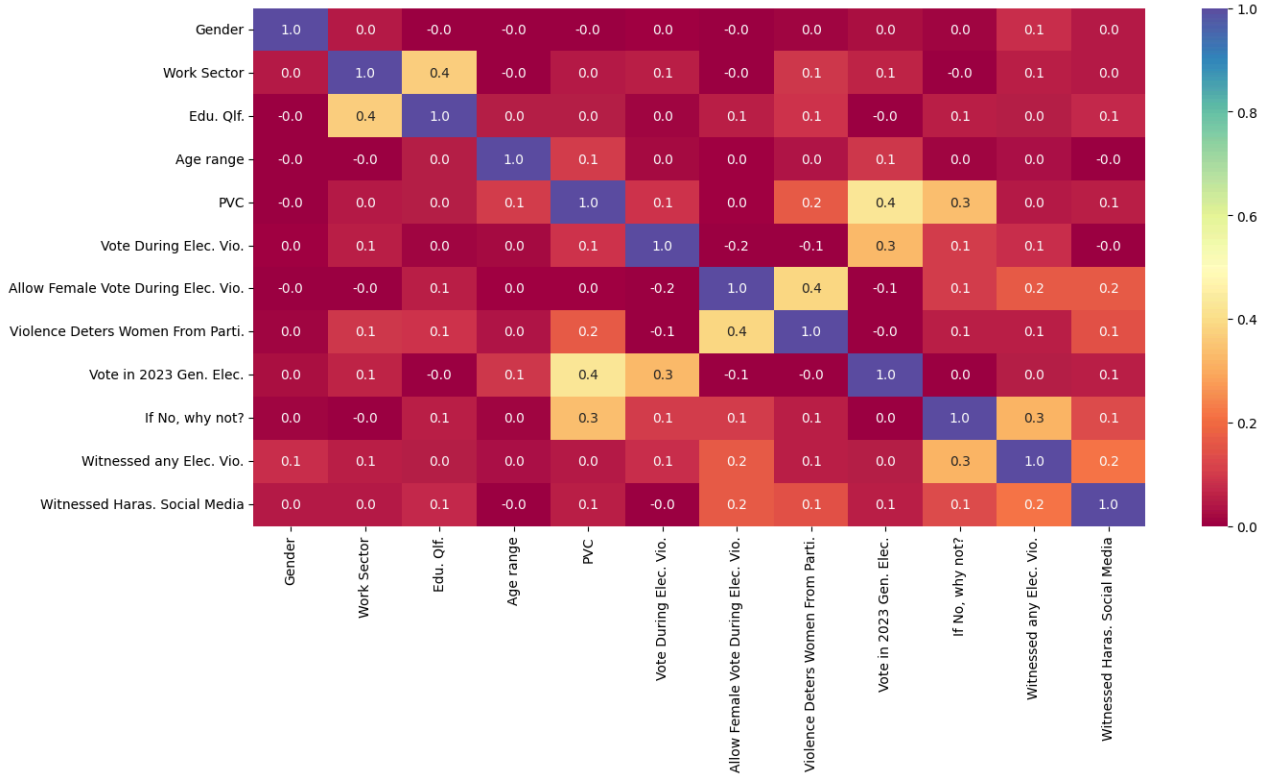
Out[73]:
```
Index(['Gender', 'Work Sector', 'Edu. Qlf.', 'Age range', 'PVC',
       'Vote During Elec. Vio.', 'Allow Female Vote During Elec. Vio.',
       'Violence Deters Women From Parti.',
       'Violence Impact Women Confi. In Parti.', 'Vote in 2023 Gen. Elec.',
       'If No, why not?', 'Witnessed any Elec. Vio.',
       'Witnessed Haras. Social Media'],
      dtype='object')
```

In [74]:
```
dataCorr.corr(method='pearson', min_periods=0, numeric_only = True)
```

Out[74]:

	Gender	Work Sector	Edu. Qlf.	Age range	PVC	Vote During Elec. Vio.	Allow Female Vote During Elec. Vio.	Violence Deters Women From Parti.	Vote in 2023 Gen. Elec.	If No, why not?	Witnessed any Elec. Vio.	Witnessed Haras. Social Media
Gender	1.000000	0.041848	-0.013503	-0.011497	-0.016611	0.004688	-0.036502	0.015561	0.025450	0.008553	0.078462	0.040072
Work Sector	0.041848	1.000000	0.370907	-0.009647	0.044513	0.056062	-0.021084	0.092999	0.062421	-0.040700	0.053180	0.040755
Edu. Qlf.	-0.013503	0.370907	1.000000	0.049578	0.047985	0.008840	0.053572	0.087862	-0.011219	0.053748	0.048310	0.068998
Age range	-0.011497	-0.009647	0.049578	1.000000	0.103473	0.017472	0.007416	0.033838	0.090162	0.009245	0.029101	-0.034038
PVC	-0.016611	0.044513	0.047985	0.103473	1.000000	0.087849	0.006357	0.162179	0.422200	0.338936	0.043147	0.053929
Vote During Elec. Vio.	0.004688	0.056062	0.008840	0.017472	0.087849	1.000000	-0.150338	-0.085752	0.322217	0.103107	0.080689	-0.014504
Allow Female Vote During Elec. Vio.	-0.036502	-0.021084	0.053572	0.007416	0.006357	-0.150338	1.000000	0.385850	-0.078540	0.103252	0.170516	0.168421
Violence Deters Women From Parti.	0.015561	0.092999	0.087862	0.033838	0.162179	-0.085752	0.385850	1.000000	-0.007867	0.057761	0.057173	0.144856
Vote in 2023 Gen. Elec.	0.025450	0.062421	-0.011219	0.090162	0.422200	0.322217	-0.078540	-0.007867	1.000000	0.001425	0.049470	0.056218
If No, why not?	0.008553	-0.040700	0.053748	0.009245	0.338936	0.103107	0.103252	0.057761	0.001425	1.000000	0.319391	0.132555
Witnessed any Elec. Vio.	0.078462	0.053180	0.048310	0.029101	0.043147	0.080689	0.170516	0.057173	0.049470	0.319391	1.000000	0.218226
Witnessed Haras. Social Media	0.040072	0.040755	0.068998	-0.034038	0.053929	-0.014504	0.168421	0.144856	0.056218	0.132555	0.218226	1.000000

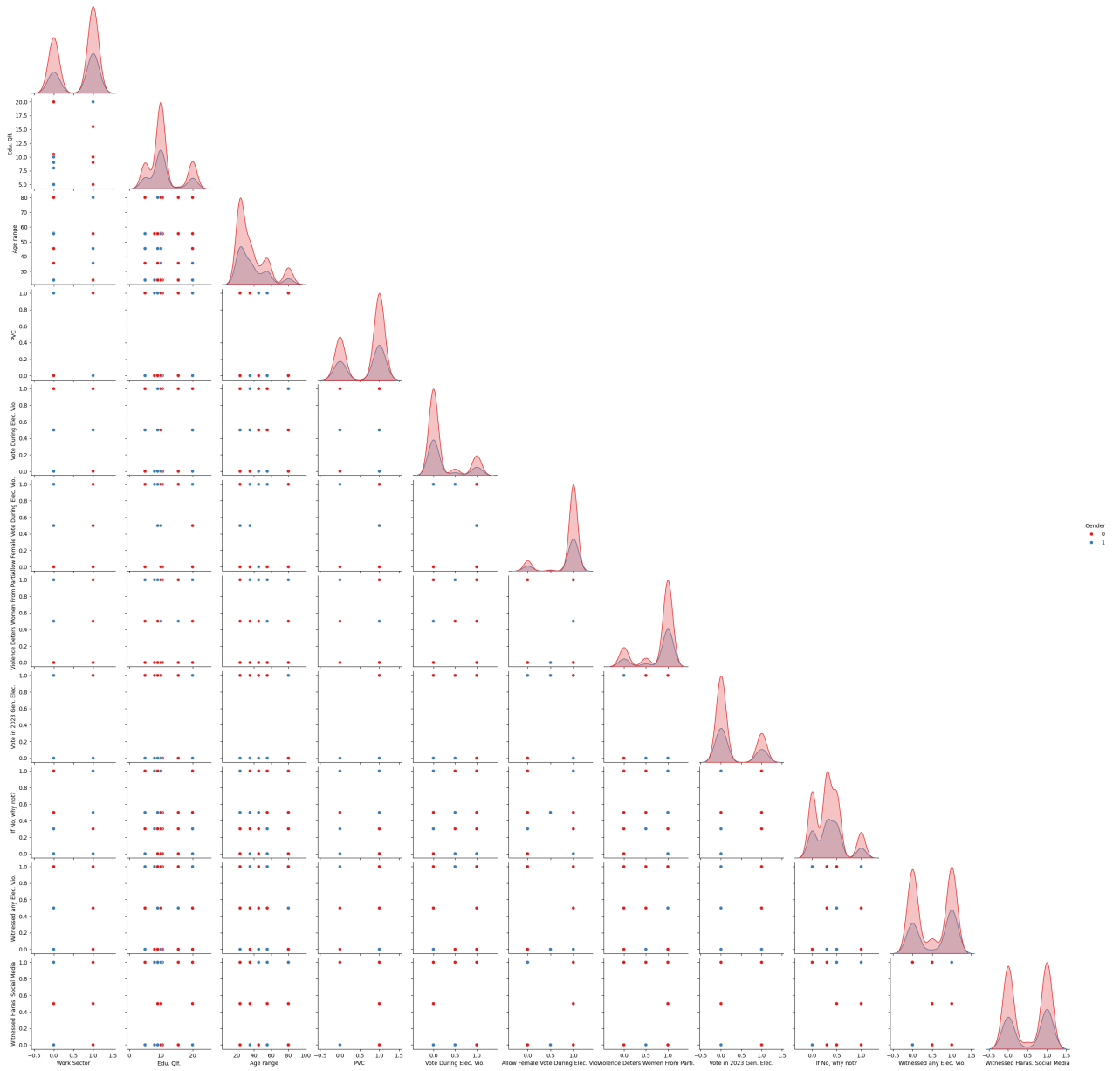
In [76]:
```
plt.figure(figsize=(15, 7))
sns.heatmap(dataCorr.corr(method='pearson'), annot=True, vmin=-0, vmax=1, fmt=".1f", cmap="Spectral")
plt.show()
```



Observation

- The education qualification is well correlated to the work sector of the respondent
- vote is 2023 election is well correlated with have a PVC and vote during election violence but slightly correlated to having witnessed harrassment on social media
- violence deter women from participating in electoral activitives is higly correlated to allowing femal to vote during electoral violence
- witnessing any form of electoral violence is also correlated to Why repondent do not vote in teh 2023 general election ("if No, why Not")
- Gender shows a slight correlation with 'witnessed any Electoral Violence'

In [165]:
```
sns.set_palette(sns.color_palette("Set1", 8))
sns.pairplot(dataCorr, hue="Gender",corner=True)
# plot_kws={'line_kws':{'color':'red'}}

plt.show()
```

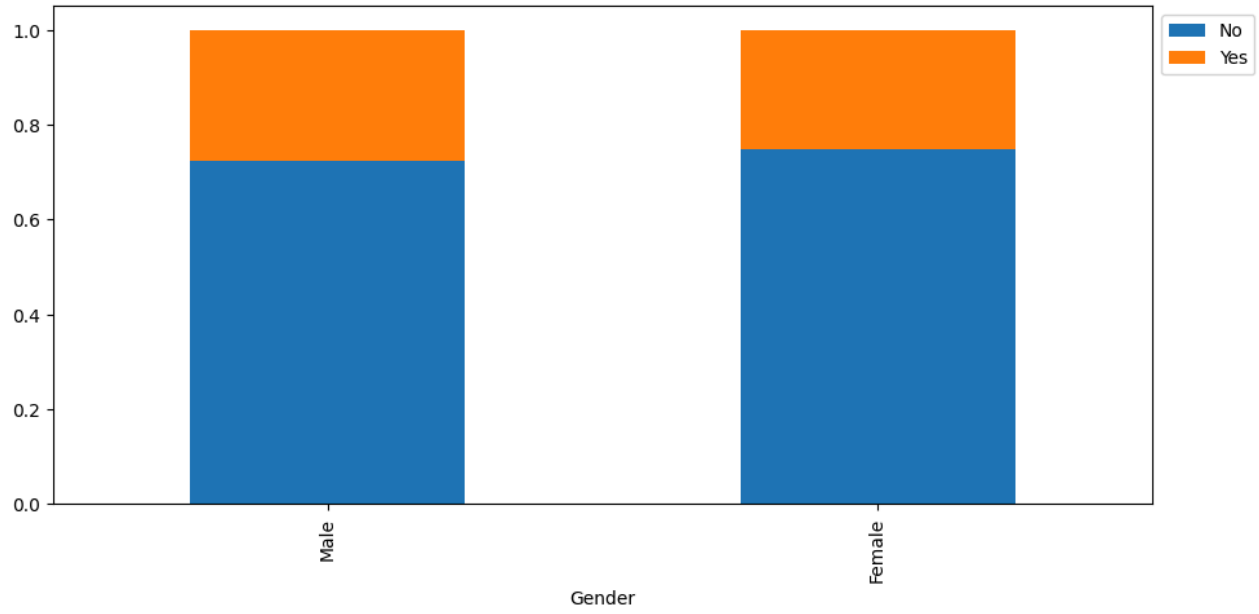


In [96]:
```
# function to plot stacked bar chart


def stacked_barplot(data, predictor, target):
    """
    Print the category counts and plot a stacked bar chart

    data: dataframe
    predictor: independent variable
    target: target variable
    """
    count = data[predictor].nunique()
    sorter = data[target].value_counts().index[-1]
    tab1 = pd.crosstab(data[predictor], data[target], margins=True).sort_values(
        by=sorter, ascending=False
    )
    print(tab1)
    print("-" * 120)
    tab = pd.crosstab(data[predictor], data[target], normalize="index").sort_values(
        by=sorter, ascending=False
    )
    tab.plot(kind="bar", stacked=True, figsize=(count + 9, 5))
    plt.legend(
        loc="lower left", frameon=False,
    )
    plt.legend(loc="upper left", bbox_to_anchor=(1, 1))
    plt.show()
```

Gender vs Did you vote in 2023 General Elections?

In [100]:
```
stacked_barplot(data, "Gender", "Did you vote in 2023 General Elections?")
```

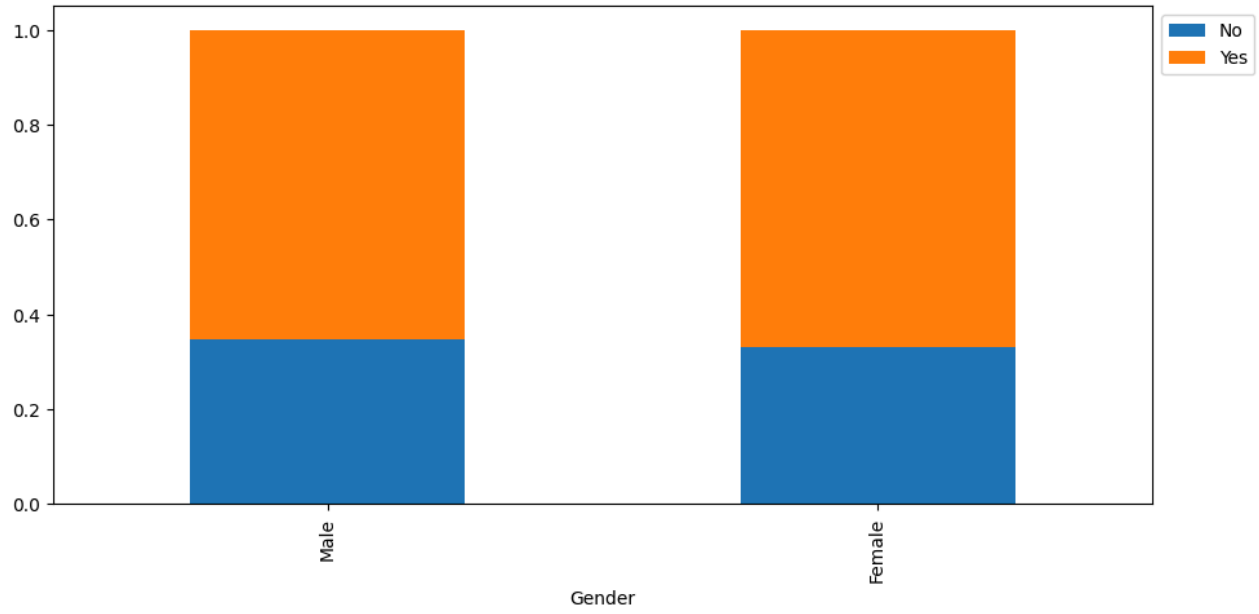
```
Did you vote in 2023 General Elections?   No  Yes  All
Gender
All                                       586  205  791
Female                                    399  134  533
Male                                      187   71  258
```



Gender vs Do you have a permanent voters card?

In [99]:
```
stacked_barplot(data, "Gender", "Do you have a permanent voters card?")
```

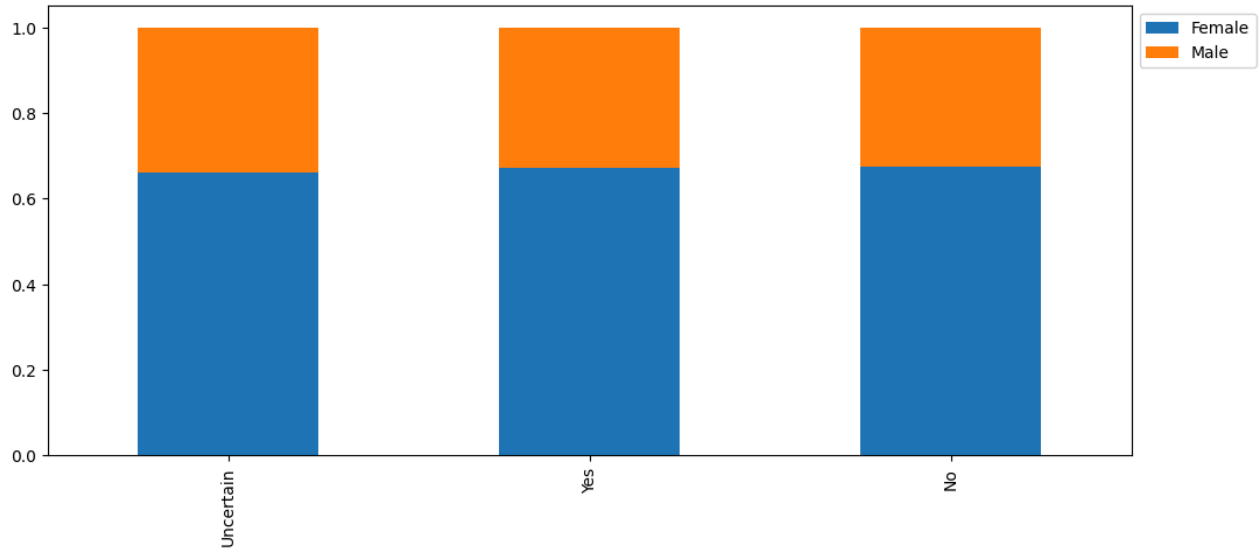
```
Do you have a permanent voters card?   No  Yes  All
Gender
All                                    267  524  791
Female                                 177  356  533
Male                                    90  168  258
```



Gender vs Are you likely to vote when there is electoral violence around you?

In [98]:
```
1 stacked_barplot(data, "Are you likely to vote when there is electoral violence around you?", "Gender")
```

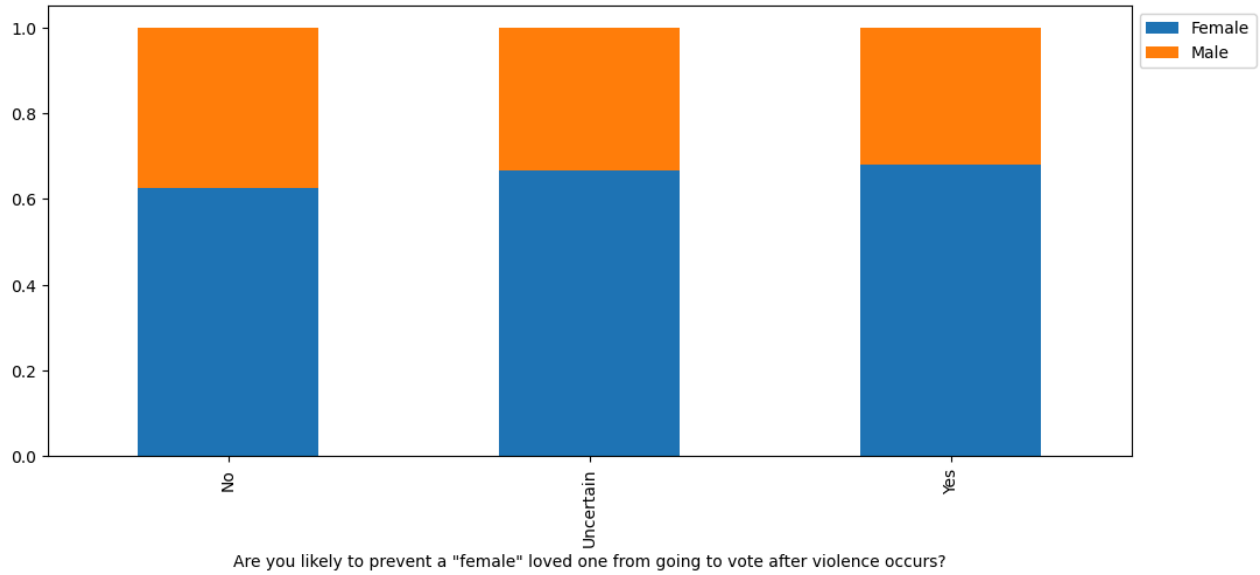
```
Gender                                          Female  Male  All
Are you likely to vote when there is electoral ...
All                                                533   258  791
No                                                 408   196  604
Yes                                                 94    46  140
Uncertain                                           31    16   47
------------------------------------------------------------------------------------------------------------------------
```



Gender vs Are you likely to prevent a "female" loved one from going to vote after violence occurs?

In [97]:
```
1 stacked_barplot(data, 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?', "Gende
```

```
Gender                                          Female  Male  All
Are you likely to prevent a "female" loved one ...
All                                                533   258  791
Yes                                                470   221  691
No                                                  57    34   91
Uncertain                                            6     3    9
------------------------------------------------------------------------------------------------------------------------
```

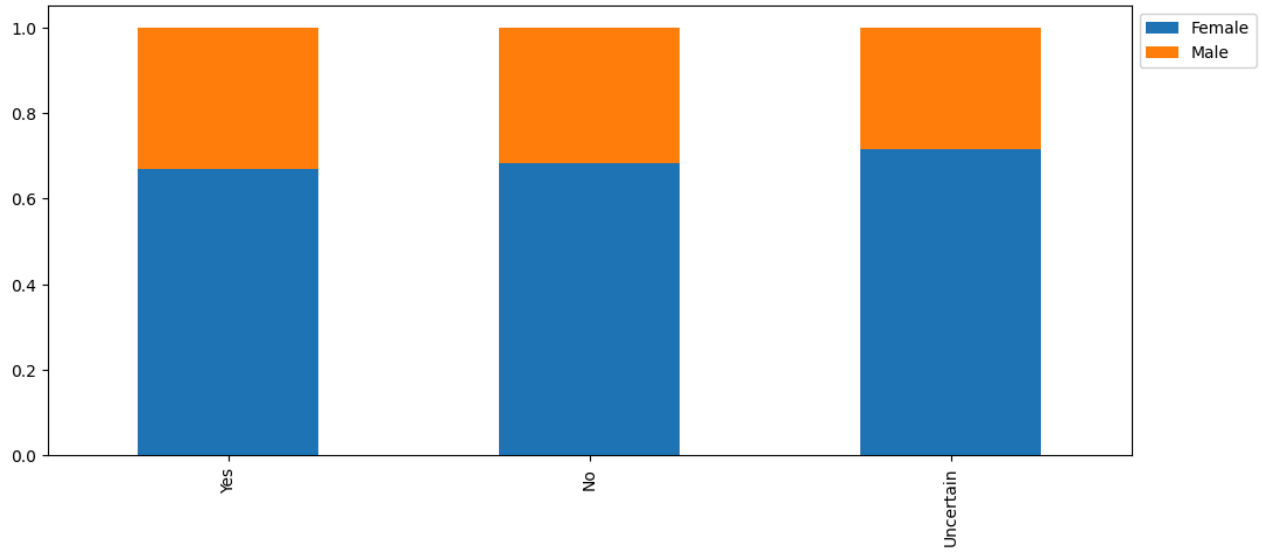


- The customers from two extreme income groups - Earning less than 40K and Earning more than 120k+ are the ones attriting the most.

Gender vs Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?

In [101]:
```
stacked_barplot(data, "Do you believe that violence deters women from participating in political activities such as rall
```

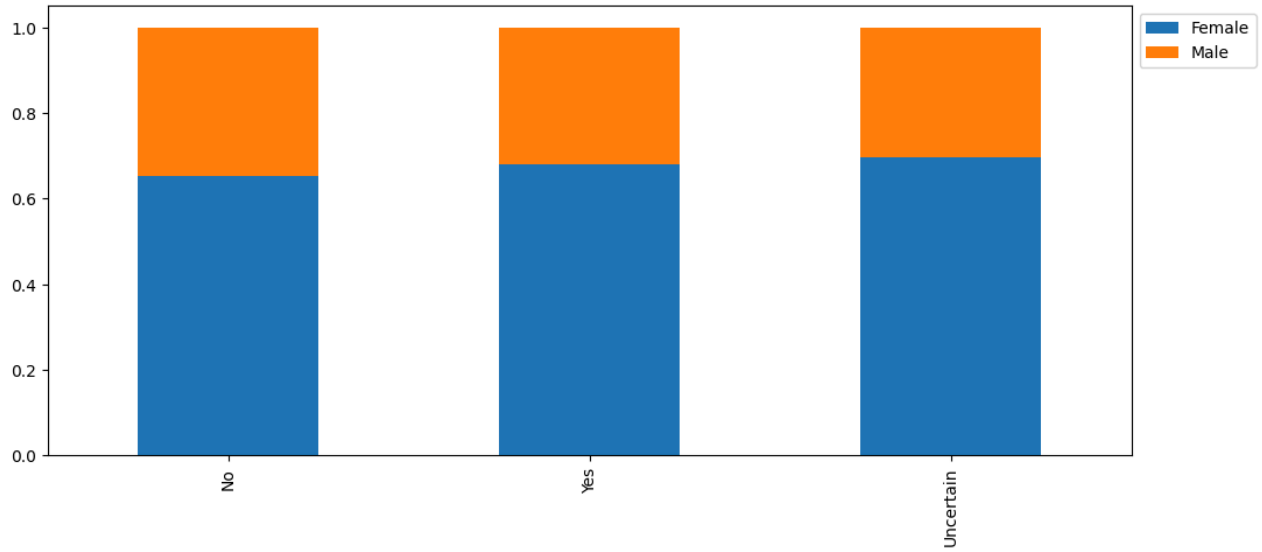
```
Gender                                              Female  Male  All
Do you believe that violence deters women from ...
All                                                    533   258  791
Yes                                                    403   200  603
No                                                      90    42  132
Uncertain                                               40    16   56
------------------------------------------------------------------------------------------------------------------------
```



Gender vs In your opinion, does violence impact the confidence of women in engaging in political activities?

In [102]:
```
stacked_barplot(data, "In your opinion, does violence impact the confidence of women in engaging in political activities
```

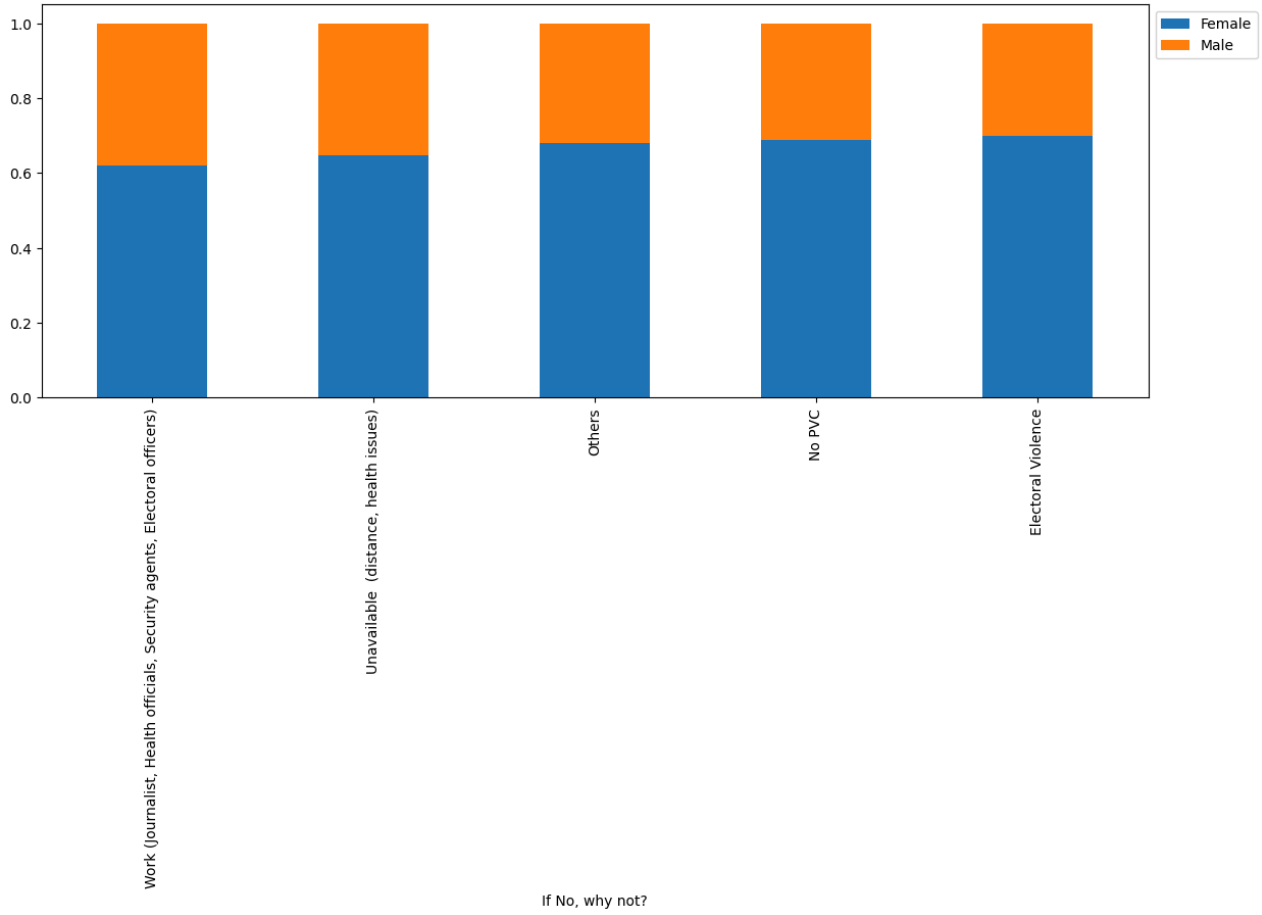
```
Gender                                              Female  Male  All
In your opinion, does violence impact the confi...
All                                                    533   258  791
Yes                                                    375   177  552
No                                                     126    67  193
Uncertain                                               32    14   46
------------------------------------------------------------------------------------------------------------------------
```



Gender vs If No, why not?

In [103]:
```
1 stacked_barplot(data, "If No, why not?", "Gender")
```

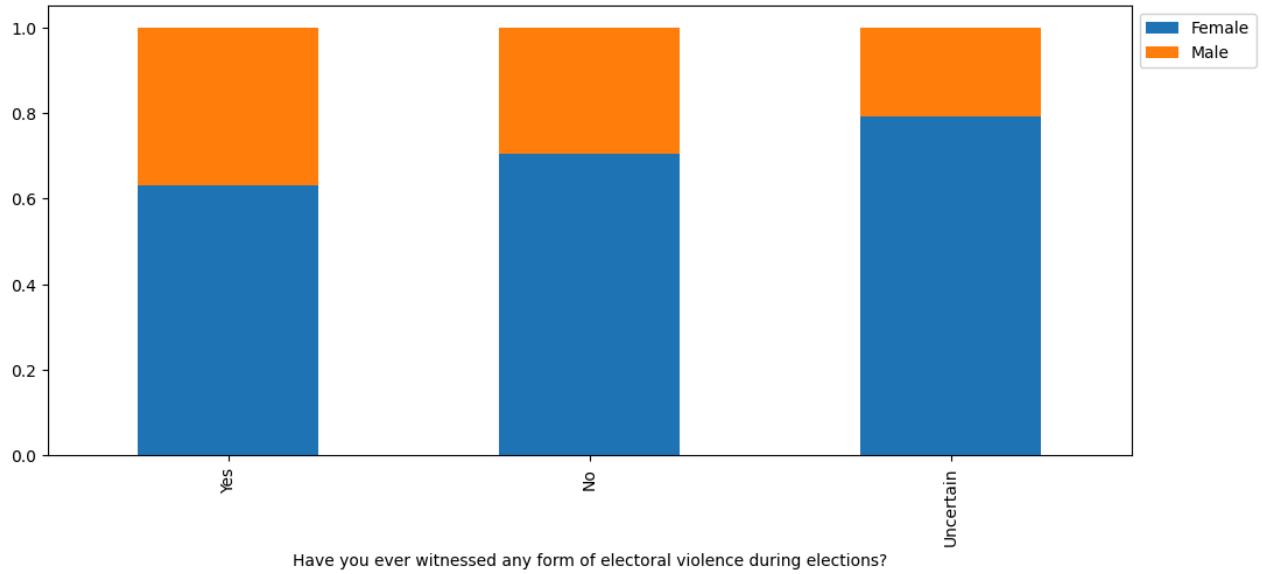
```
Gender                                          Female  Male  All
If No, why not?
All                                                533   258  791
Others                                             186    87  273
No PVC                                             150    68  218
Unavailable  (distance, health issues)             103    56  159
Electoral Violence                                  58    25   83
Work (Journalist, Health officials, Security ag...  36    22   58
------------------------------------------------------------------------------------------------------------------------
```



Gender vs Have you ever witnessed any form of electoral violence during elections?

In [104]:
```
1 stacked_barplot(data, "Have you ever witnessed any form of electoral violence during elections?", "Gender")
```

```
Gender                                              Female  Male  All
Have you ever witnessed any form of electoral v...
All                                                    533   258  791
Yes                                                    249   146  395
No                                                     242   101  343
Uncertain                                               42    11   53
------------------------------------------------------------------------------------------------------------------------
```



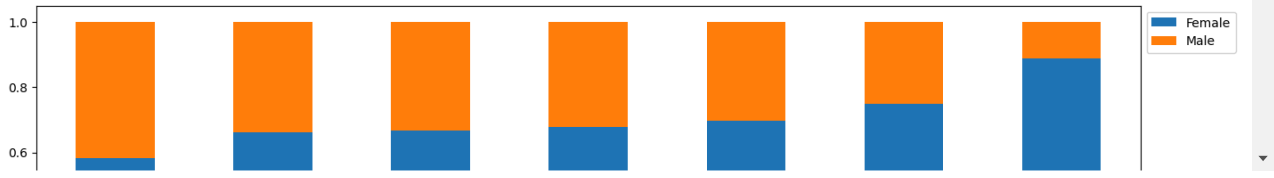
Gender vs Have you ever witnessed any form of harassment on social media?

In [105]:
```
1 stacked_barplot(data, "Have you ever witnessed any form of harassment on social media?", "Gender")
```

```
Gender                                              Female  Male  All
Have you ever witnessed any form of harassment ...
All                                                    533   258  791
Yes                                                    262   140  402
No                                                     251   113  364
Uncertain                                               20     5   25
------------------------------------------------------------------------------------------------------------------------
```

In [106]:
```
1 stacked_barplot(data, "Educational Qualification", "Gender")
```

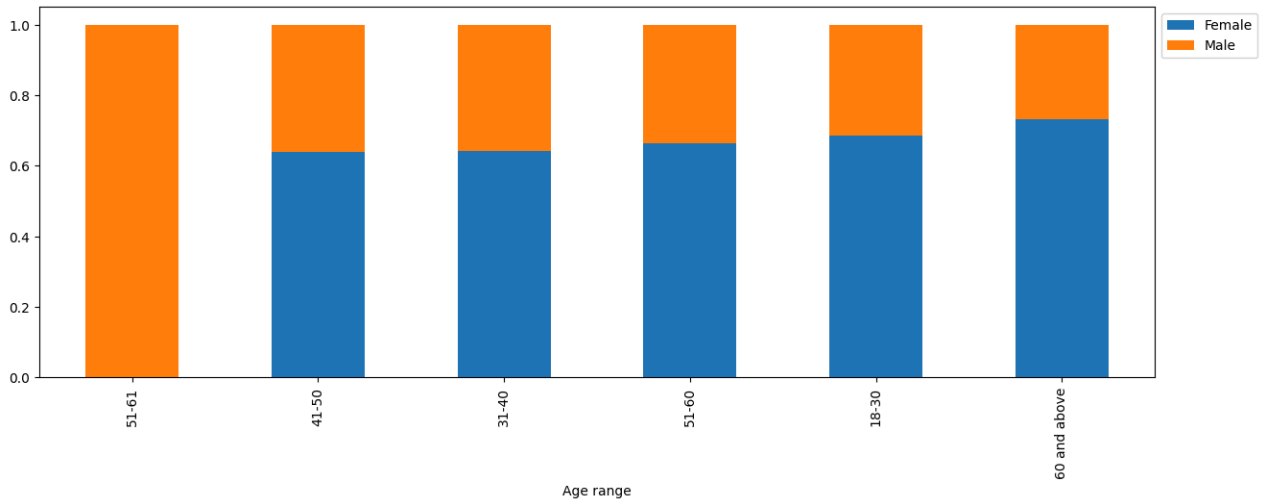
```
Gender                    Female  Male  All
Educational Qualification
All                          533   258  791
HND, B.Sc.                   290   148  438
SSCE and below                95    45  140
Postgraduate                 101    44  145
ND, NCE                       26    13   39
Mbbs in view                   7     5   12
Undergraduate                  6     2    8
Btech                          8     1    9
------------------------------------------------------------------------------------------------------------------------
```



Gender vs Age range

In [107]:
```
stacked_barplot(data, "Age range", "Gender")
```

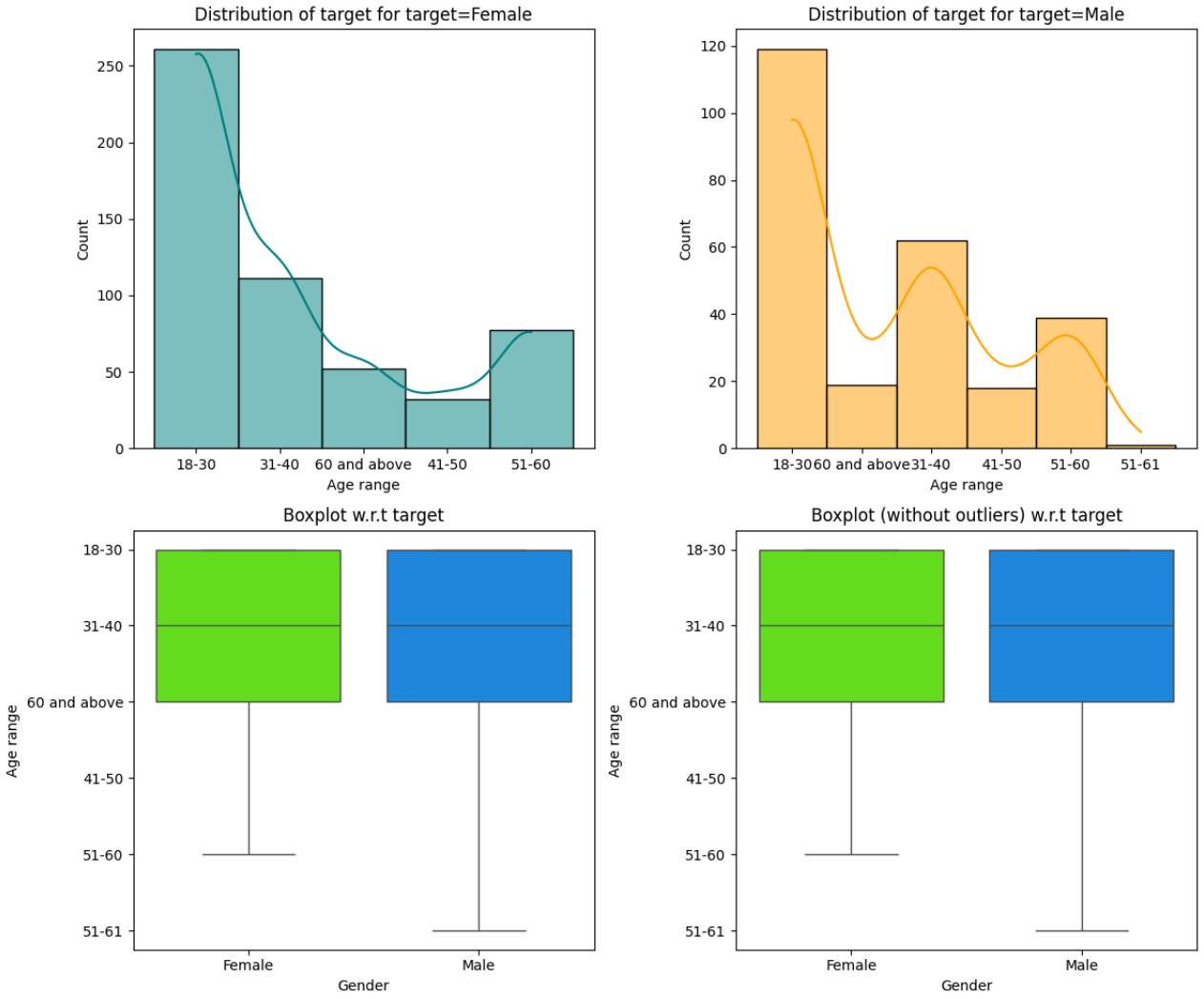
```
Gender        Female  Male  All
Age range
All              533   258  791
18-30            261   119  380
31-40            111    62  173
51-60             77    39  116
60 and above      52    19   71
41-50             32    18   50
51-61              0     1    1
```



In [108]:
```
### Function to plot distributions


def distribution_plot_wrt_target(data, predictor, target):

    fig, axs = plt.subplots(2, 2, figsize=(12, 10))

    target_uniq = data[target].unique()

    axs[0, 0].set_title("Distribution of target for target=" + str(target_uniq[0]))
    sns.histplot(
        data=data[data[target] == target_uniq[0]],
        x=predictor,
        kde=True,
        ax=axs[0, 0],
        color="teal",
    )

    axs[0, 1].set_title("Distribution of target for target=" + str(target_uniq[1]))
    sns.histplot(
        data=data[data[target] == target_uniq[1]],
        x=predictor,
        kde=True,
        ax=axs[0, 1],
        color="orange",
    )

    axs[1, 0].set_title("Boxplot w.r.t target")
    sns.boxplot(data=data, x=target, y=predictor, ax=axs[1, 0], palette="gist_rainbow")

    axs[1, 1].set_title("Boxplot (without outliers) w.r.t target")
    sns.boxplot(
        data=data,
        x=target,
        y=predictor,
        ax=axs[1, 1],
        showfliers=False,
        palette="gist_rainbow",
    )

    plt.tight_layout()
    plt.show()
```

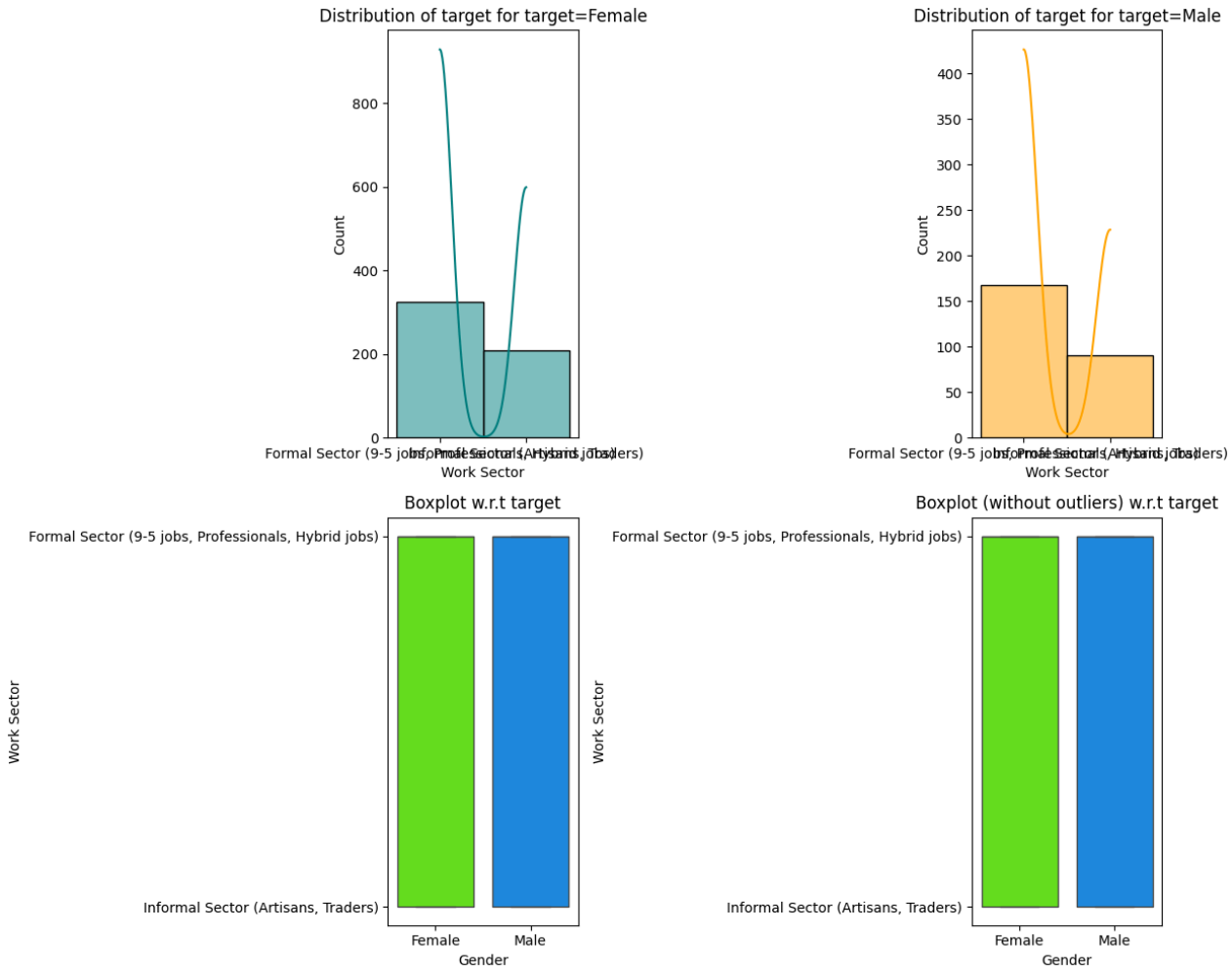
Gender vs Age range

In [109]:
```
1 distribution_plot_wrt_target(data, "Age range", "Gender")
```



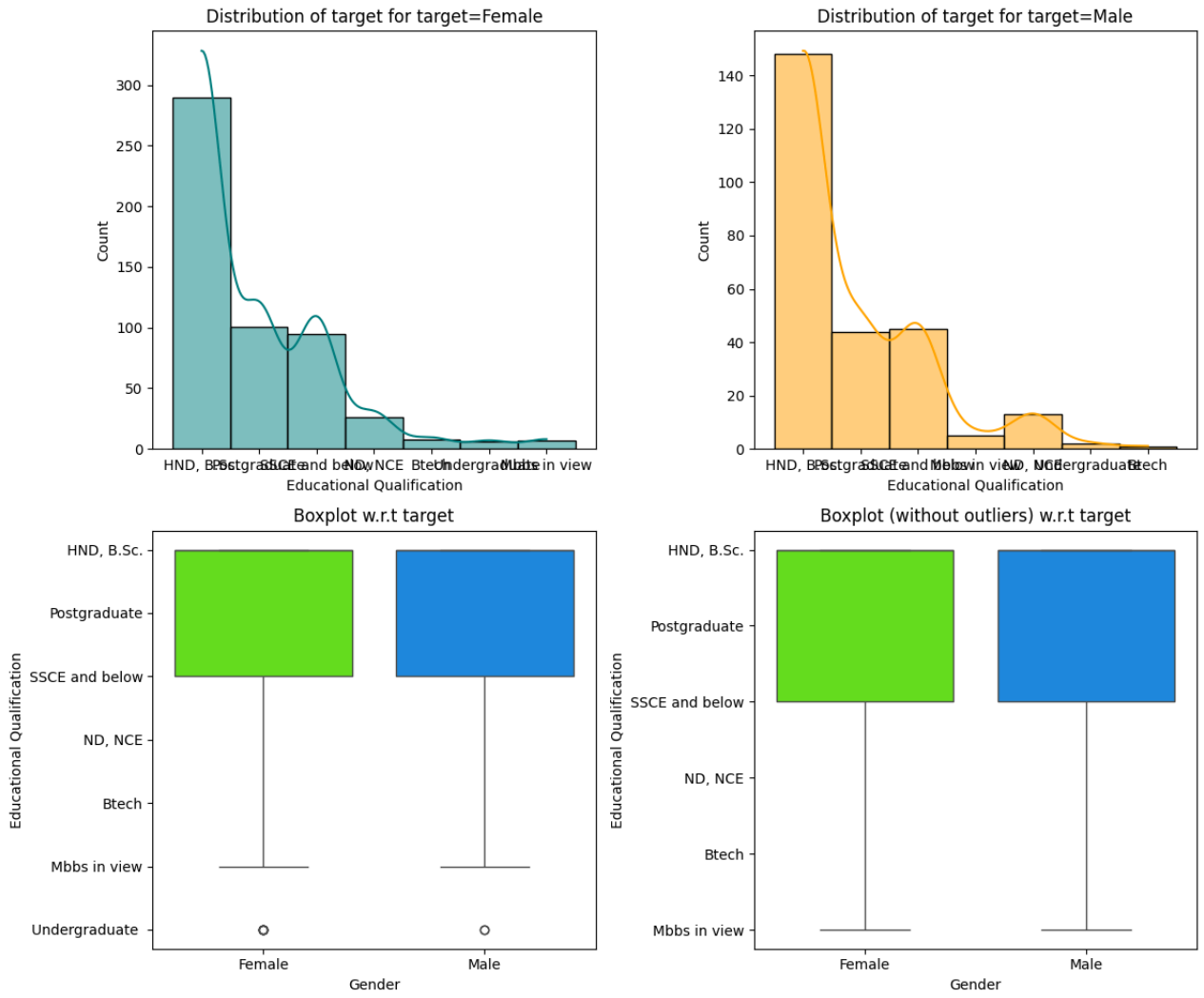
Gender vs Work Sector

In [110]:
```
1 distribution_plot_wrt_target(data, "Work Sector", "Gender")
```



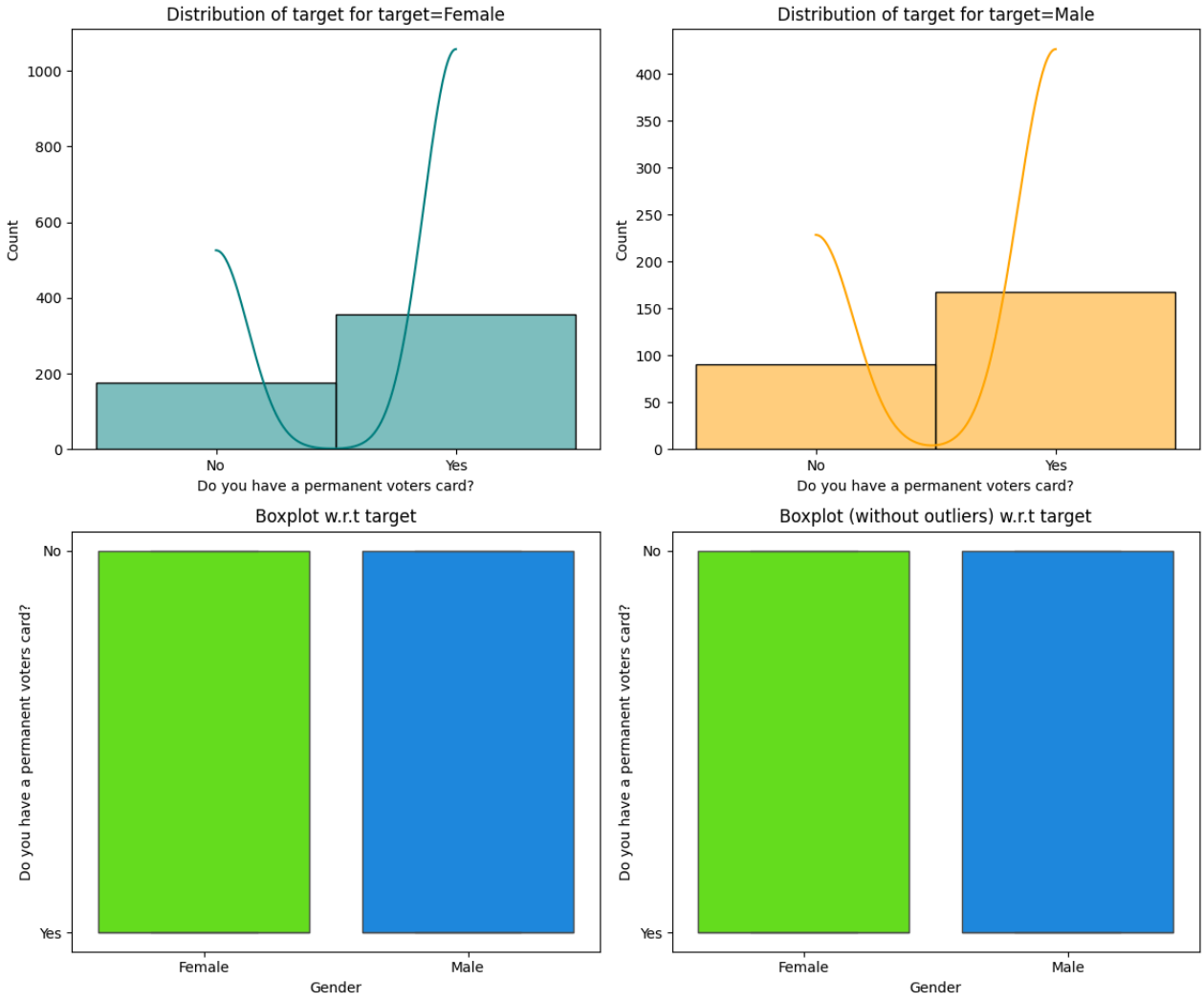
Gender vs Educational Qualification

In [111]:
```
1 distribution_plot_wrt_target(data, "Educational Qualification", "Gender")
```



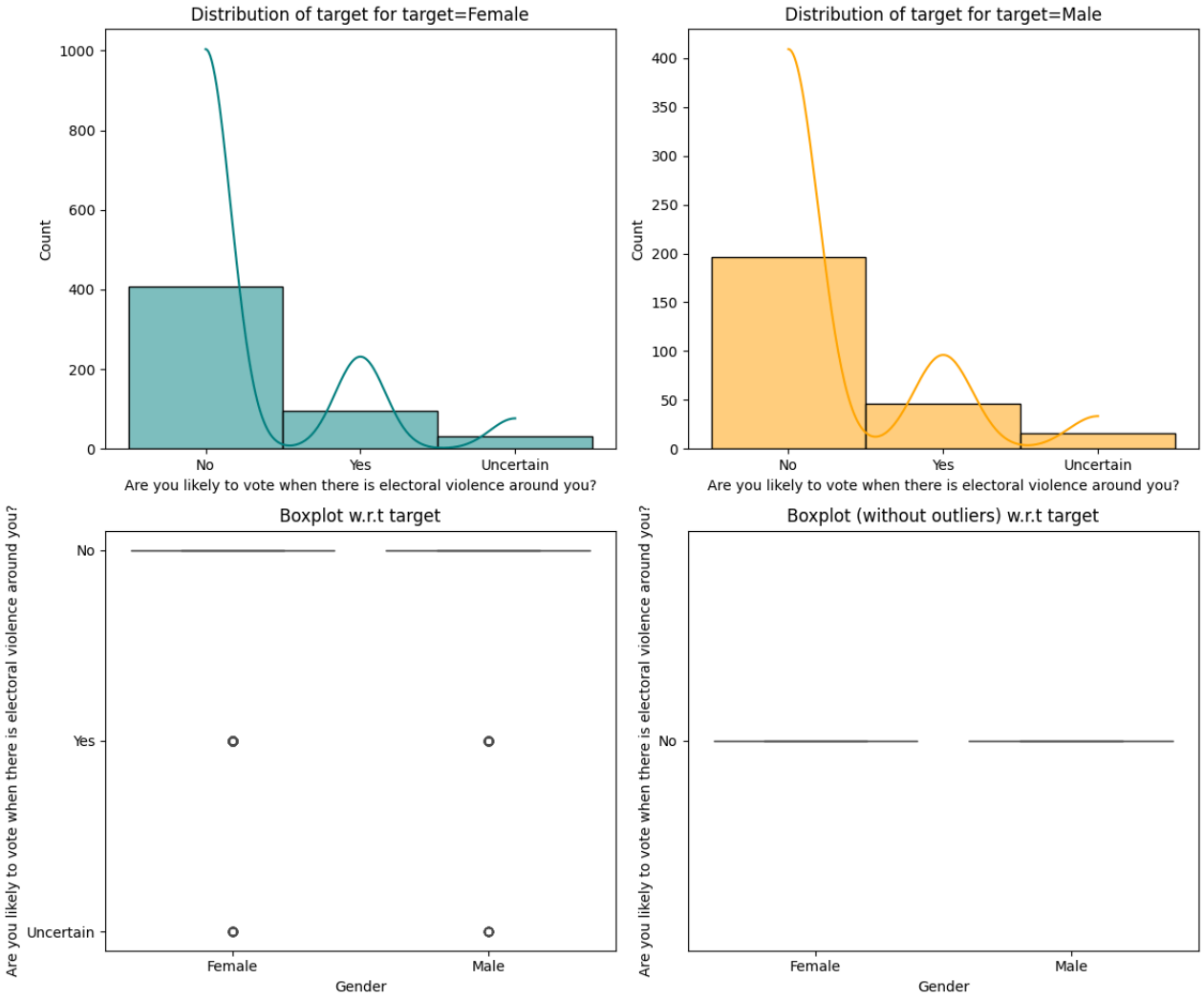
Gender vs Do you have a permanent voters card?

In [113]:
```
1 distribution_plot_wrt_target(data, "Do you have a permanent voters card?", "Gender")
```



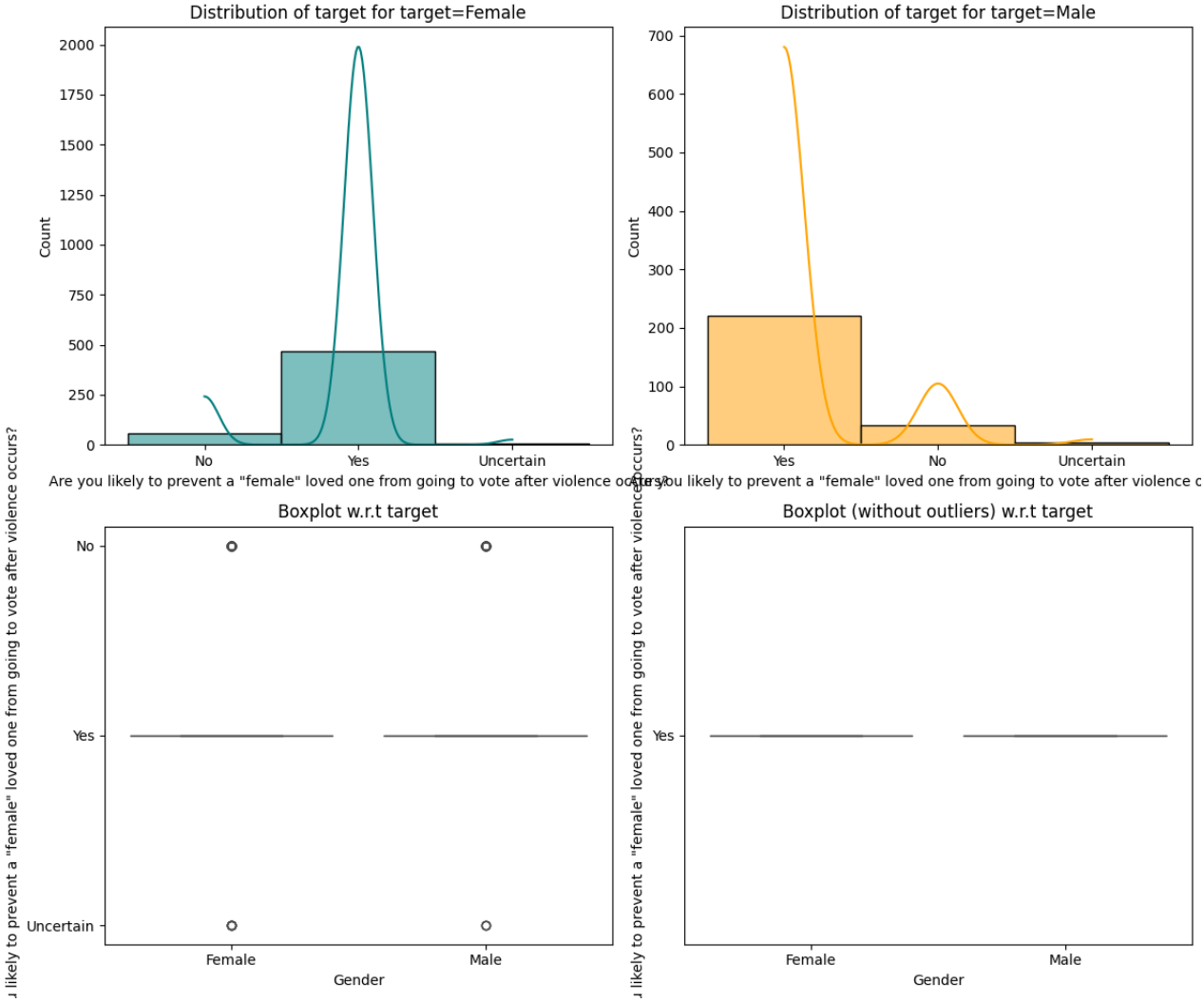
Gender vs Are you likely to vote when there is electoral violence around you?

In [114]:
```
distribution_plot_wrt_target(data, "Are you likely to vote when there is electoral violence around you?", "Gender")
```



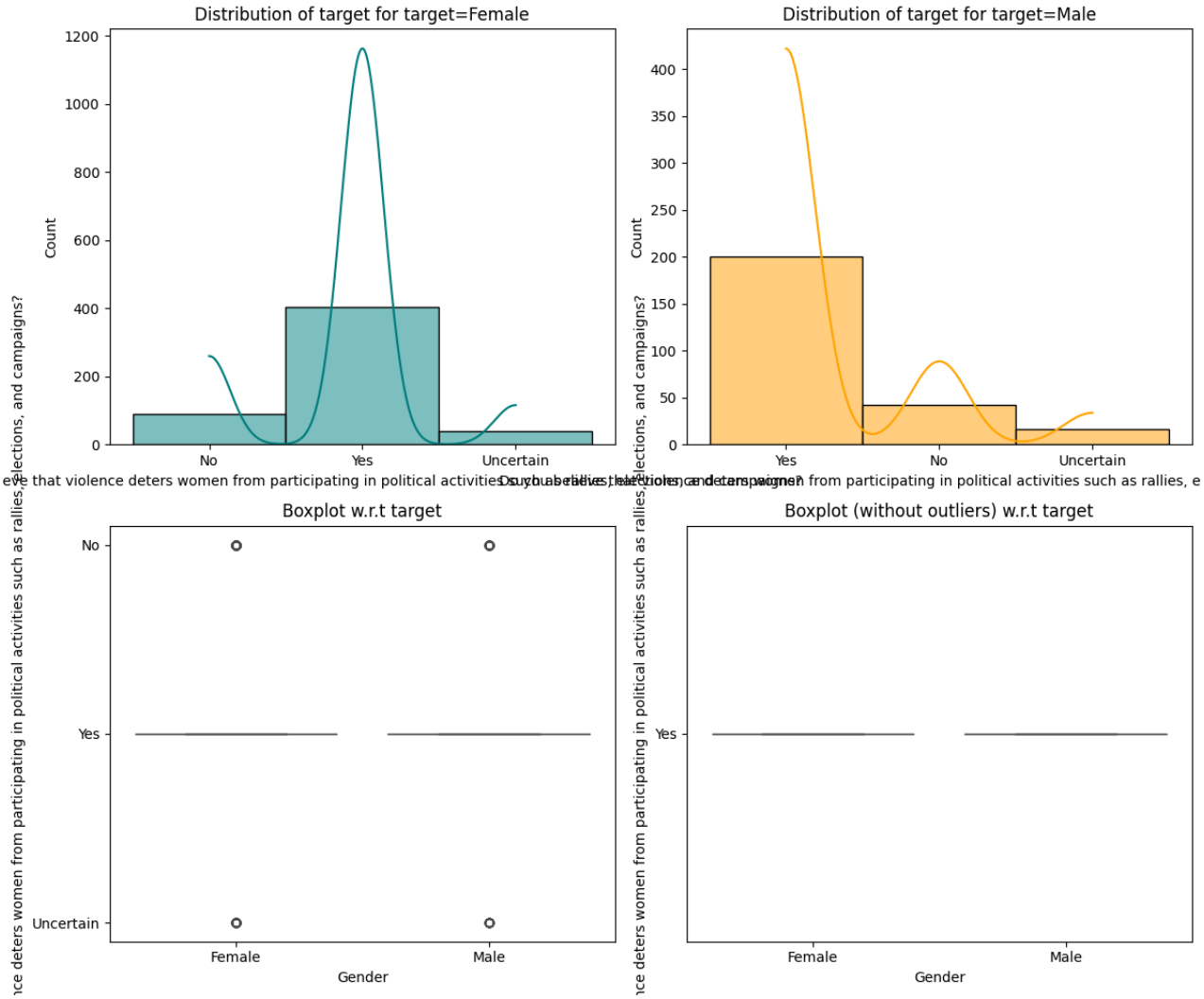
Gender vs Are you likely to prevent a "female" loved one from going to vote after violence occurs?

In [116]:
```
distribution_plot_wrt_target(data,  'Are you likely to prevent a "female" loved one from going to vote after violence oc
```



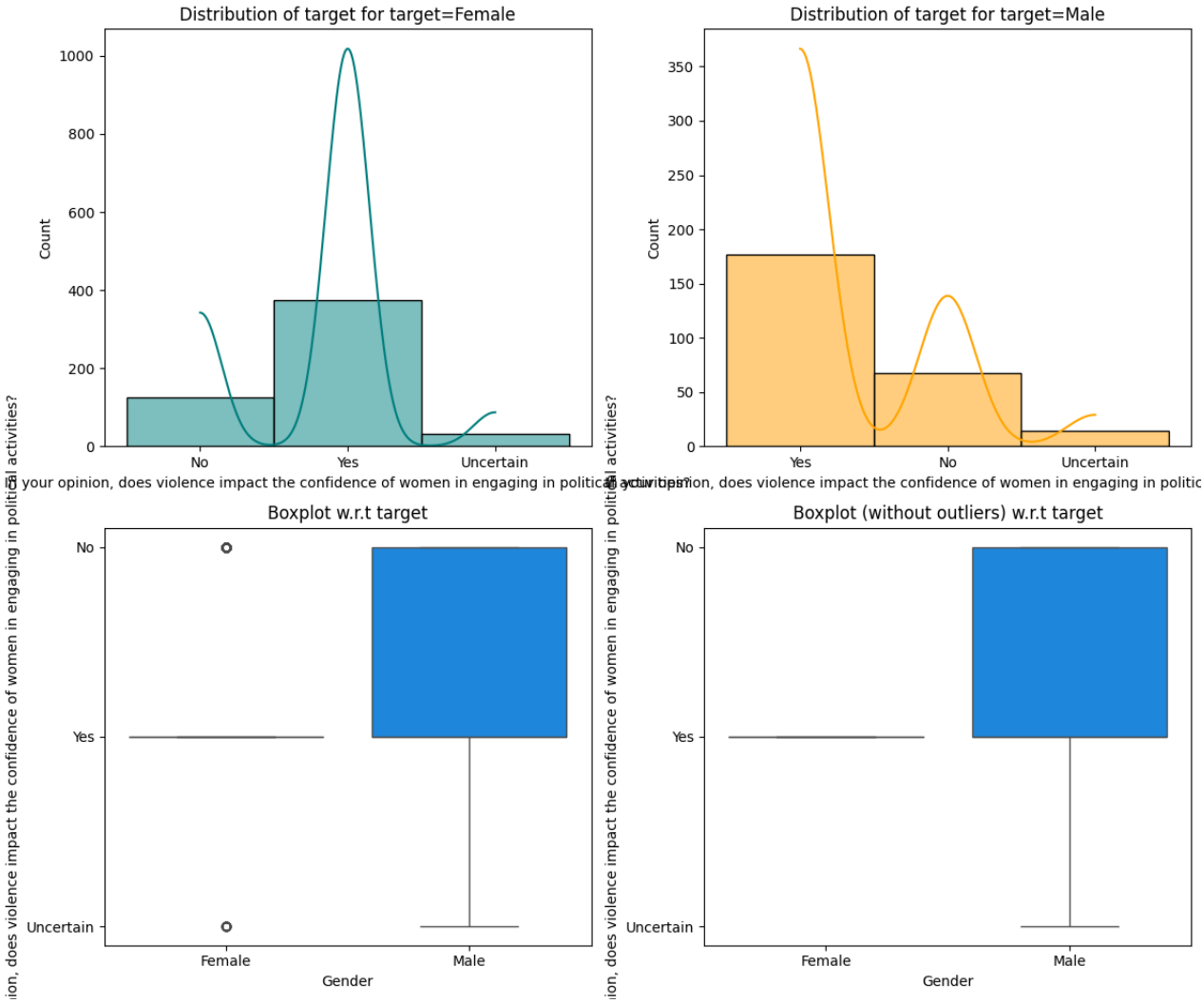
Gender vs Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaigns?

In [117]:
```
distribution_plot_wrt_target(data, "Do you believe that violence deters women from participating in political activities
```



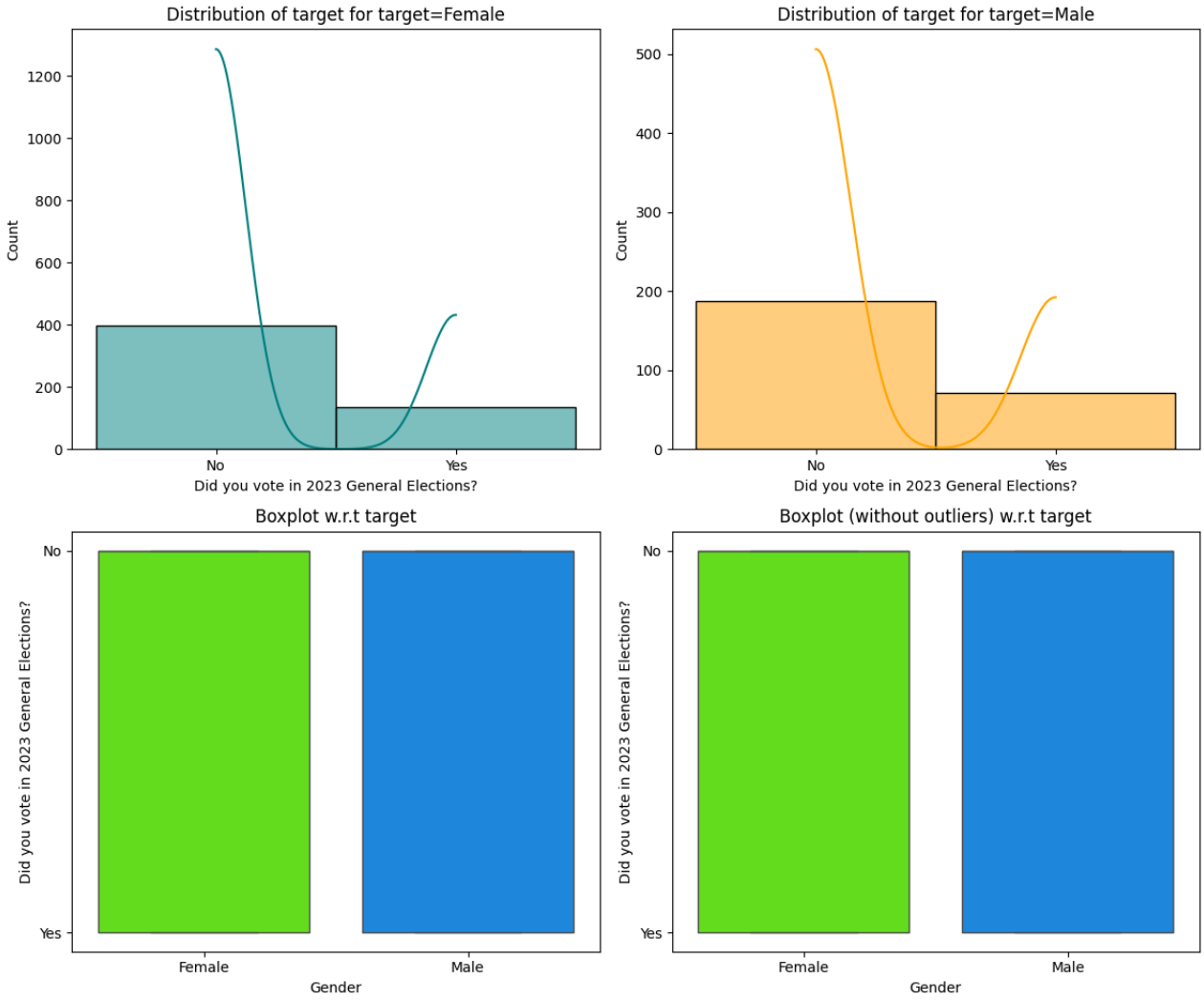
Gender vs In your opinion, does violence impact the confidence of women in engaging in political activities?

In [119]:
```
distribution_plot_wrt_target(data, "In your opinion, does violence impact the confidence of women in engaging in politic
```



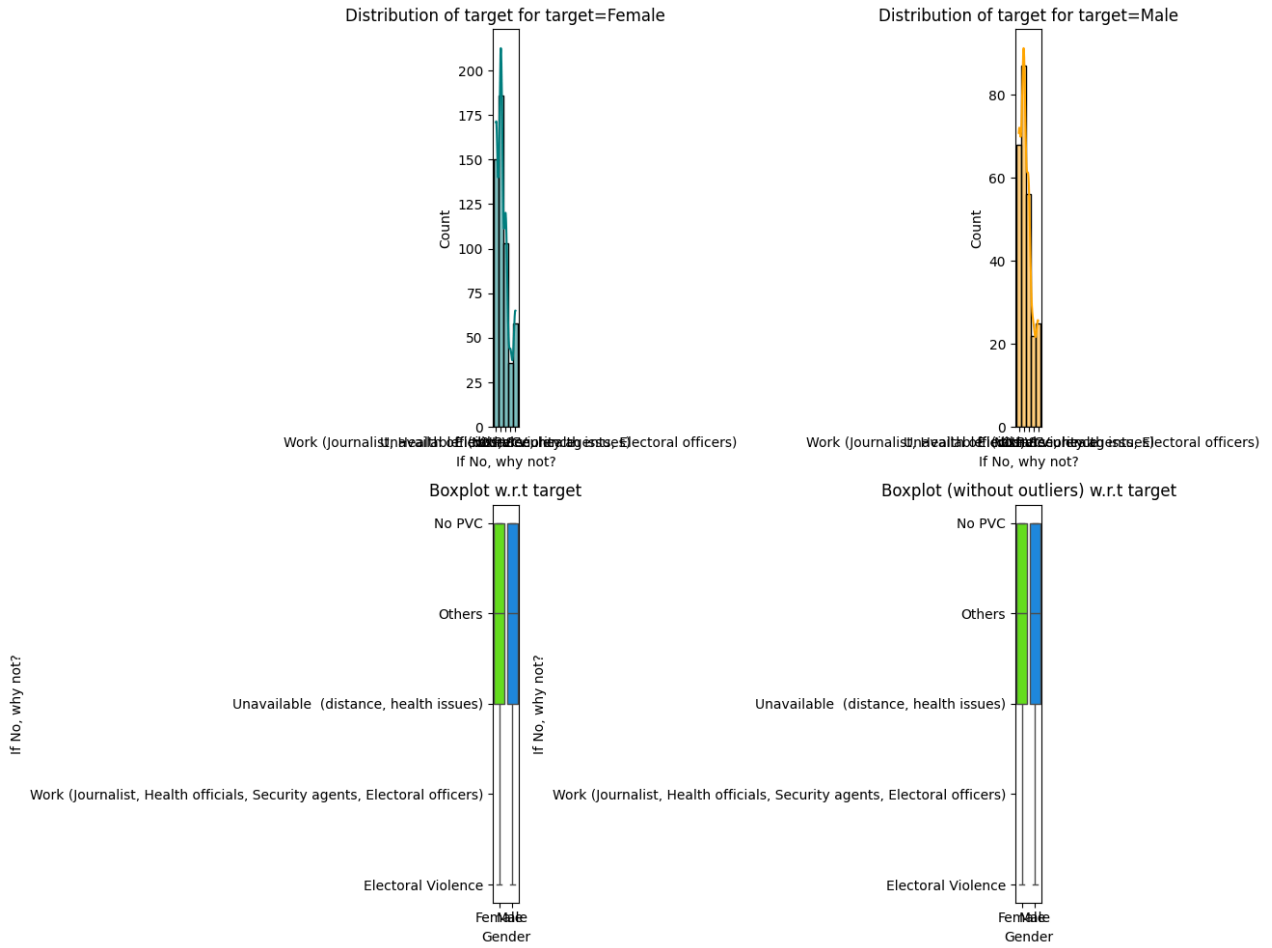
Gender vs Did you vote in 2023 General Elections?

In [120]:
```
distribution_plot_wrt_target(data, "Did you vote in 2023 General Elections?", "Gender")
```

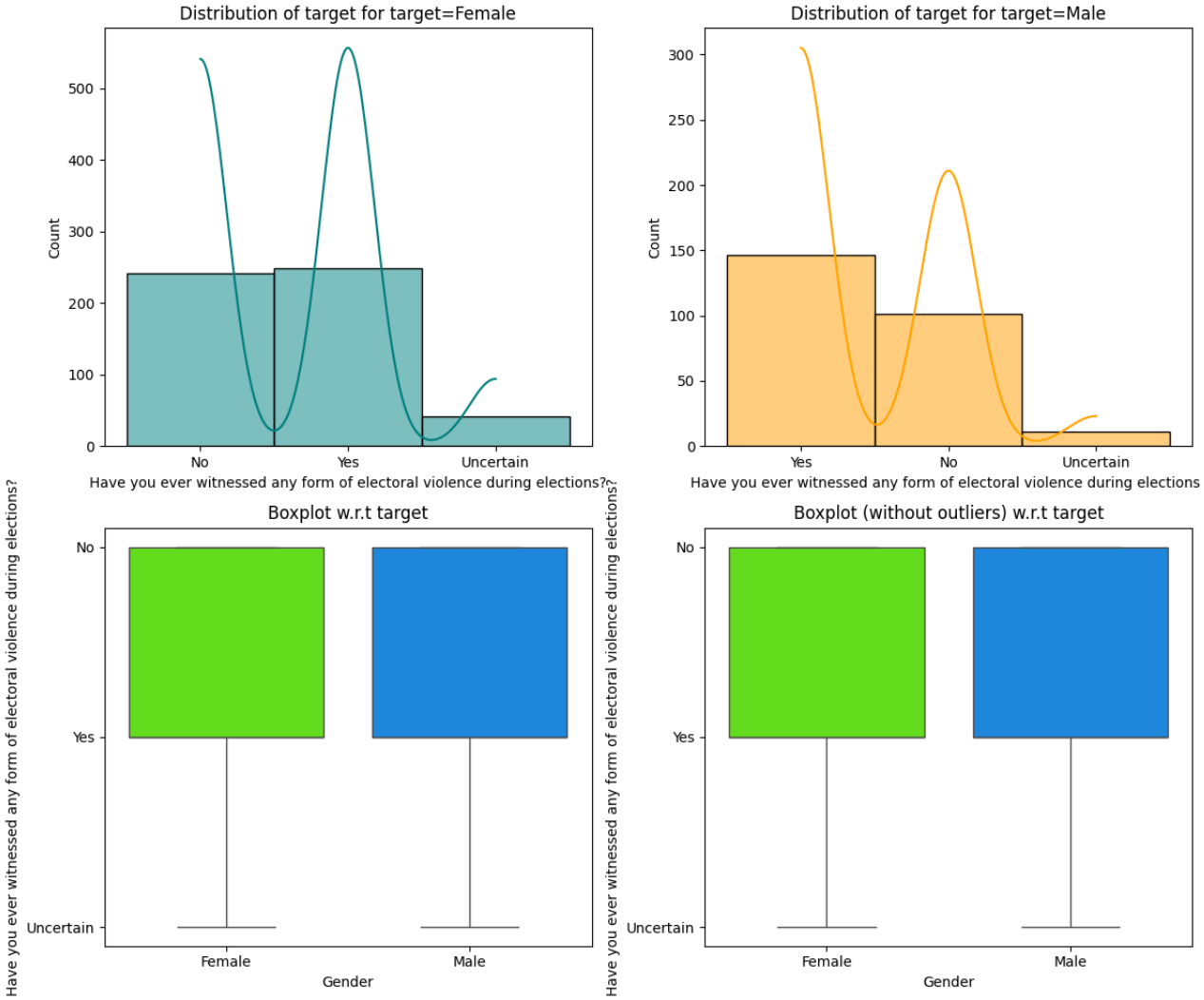


Gender vs If No, why not?

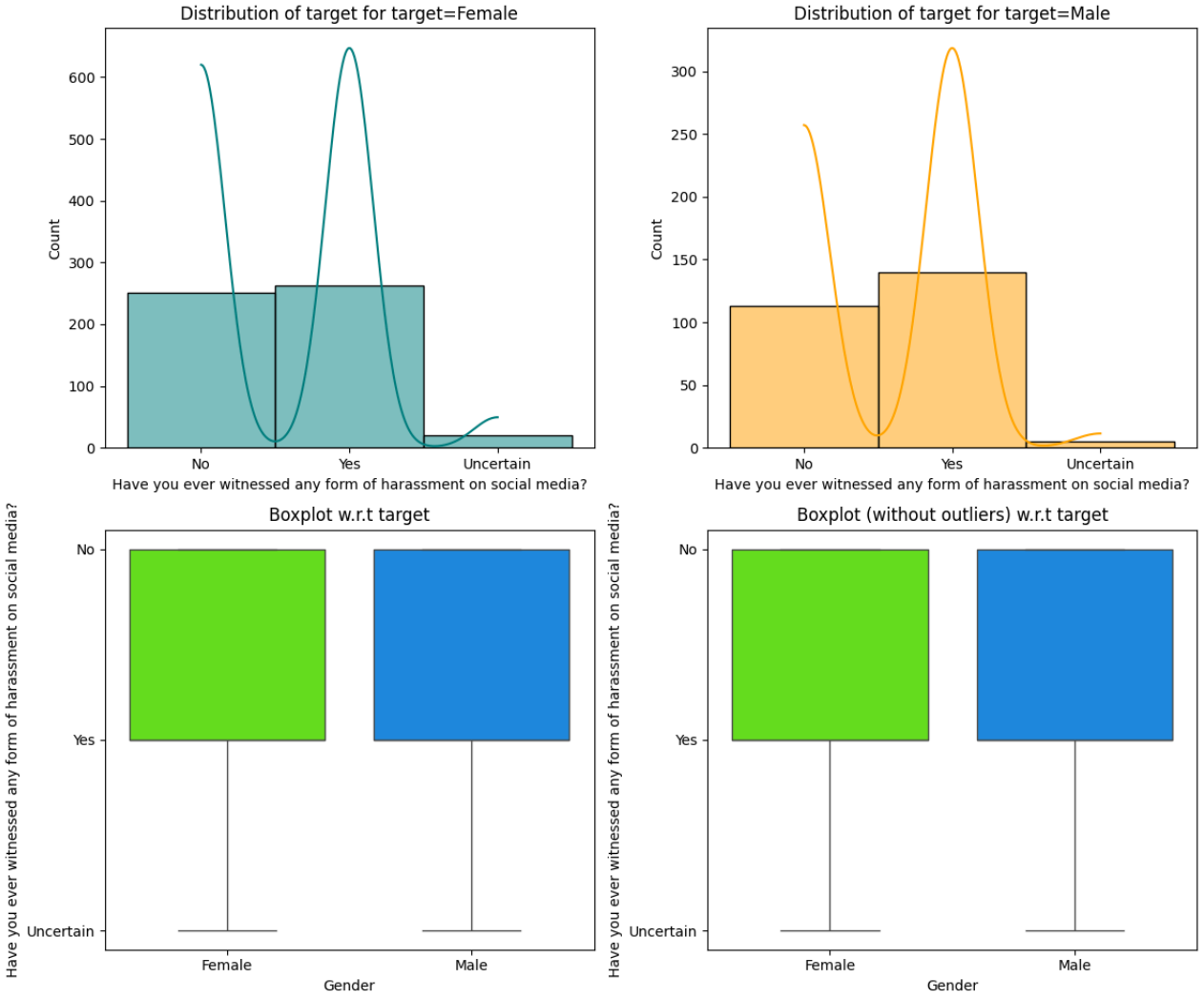
In [123]:
```
1 distribution_plot_wrt_target(data, "If No, why not?", "Gender")
```



In [125]:
```
distribution_plot_wrt_target(data, "Have you ever witnessed any form of electoral violence during elections?", "Gender")
```



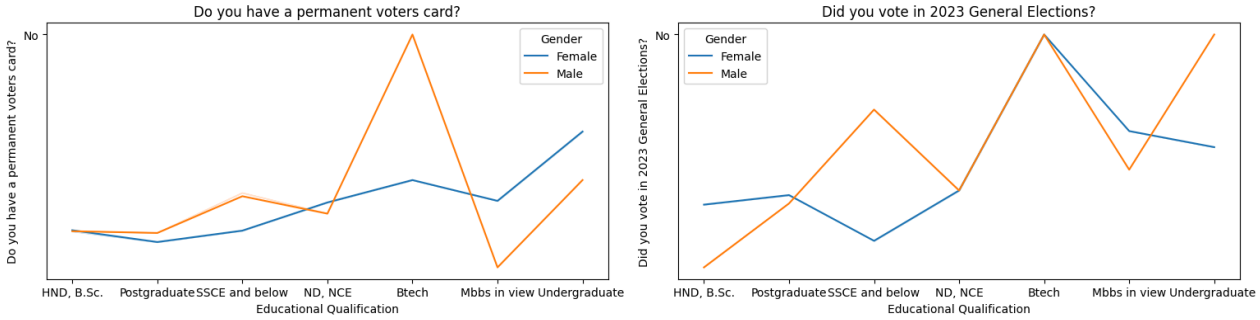
In [126]:
```
distribution_plot_wrt_target(data, "Have you ever witnessed any form of harassment on social media?", "Gender")
```



Multivariant Analysis

In [145]:
```
def line_plot(cols):
    plt.figure(figsize=(23, 11))
    for i, variable in enumerate(cols):
        plt.subplot(3, 3, i + 1)
        sns.lineplot(data=data, x="Educational Qualification", y=variable, hue="Gender", ci=0)
        plt.tight_layout()
        plt.title(variable)
    plt.show()
```

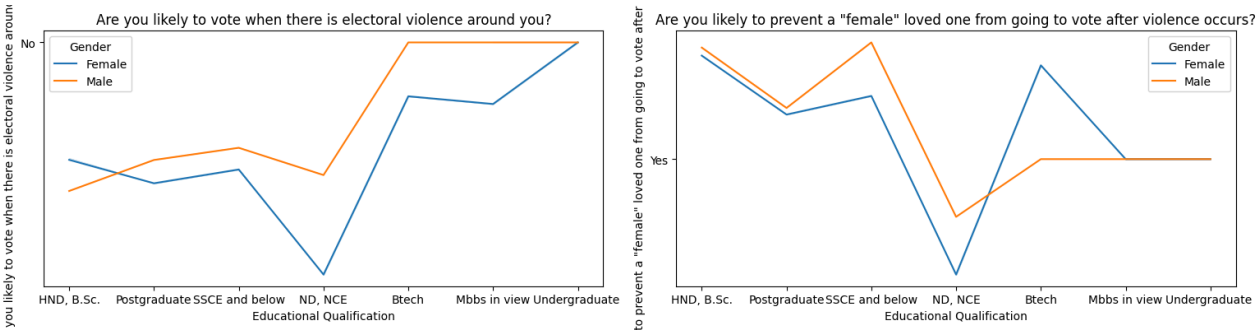
In [146]:
```
cols = data[["Do you have a permanent voters card?","Did you vote in 2023 General Elections?"]].columns.tolist()
line_plot(cols)
```



Observation

- Majority If the females that do not vot in the 2023 General Election hold BTech follow by MBBS in View
- Majority of the female that have PVC are Undergraduate Studnet follow by BTech and Mbbs in view Holder

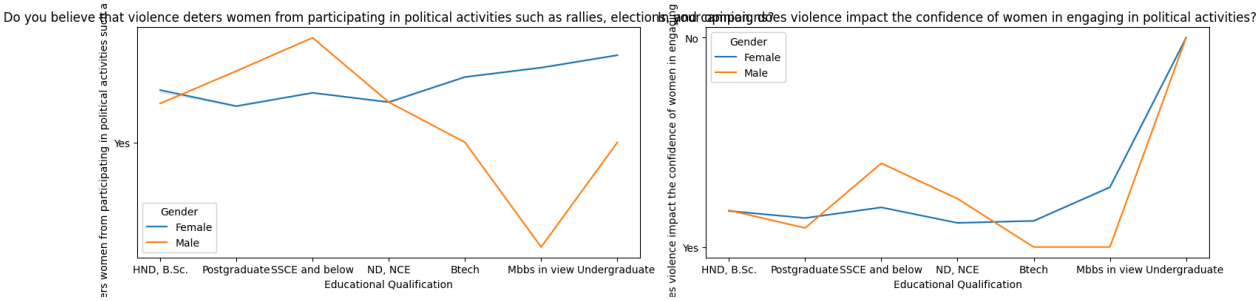
In [147]:
```
re is electoral violence around you?','Are you likely to prevent a "female" loved one from going to vote after violence occur
```



Observation

- Majority of the females that their education level is undergarduate indicate they are not likely to vote when there is electoral violence around them
- Although more male shows that are not likely to vote when there is electoral violence around them
- Females with ND,NCE qualification still show willingness to vote despicte electoral violence around them
- only Females with ND,NCE are uncertain to prevent a "female" loved one's from going to vote after violence occurs

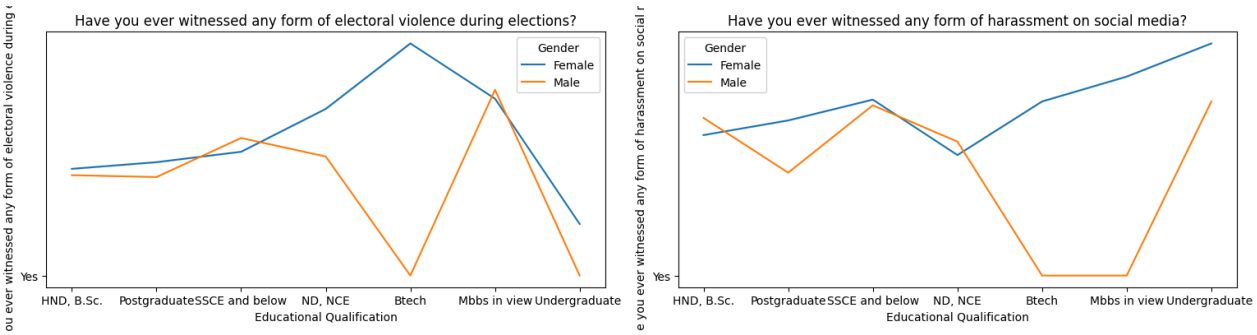
In [148]:
```
data[
Do you believe that violence deters women from participating in political activities such as rallies, elections, and campaign
In your opinion, does violence impact the confidence of women in engaging in political activities?']
mns.tolist()
lot(cols)
```



Observation

- Male respondent who have Mbbs inn view are **uncertain if** violence deters women from participating in pratical activities sucg as as rallies, elections, and campaigns
- Male respondent who have SSCE and Below claims that violence **does not** deters women from participating in pratical activities sucg as as rallies, elections, and campaigns
- All the female respondent, irrespective of their educational level claims that **agree** that violence deters women from participating in pratical activities sucg as as rallies, elections, and campaigns

In [149]:
```
cols = data[
    ["Have you ever witnessed any form of electoral violence during elections?",
     'Have you ever witnessed any form of harassment on social media?']
].columns.tolist()
line_plot(cols)
```



Observation

- Male respondent who have BTech and undergraduate are are the lowest among all the respondent that has ever witnessed any form of electoral violence during elections
- All the female respondent, irrespective of their educational level claims that **agree** that they have witnessed any form of electoral violence during elections with females with BTech beign the highest.

- Male respondent who have BTech and Mbbs in view are are the lowest among all the respondent that has ever witnessed any form of harrassment on social media
- All the female respondent, irrespective of their educational level claims that **agree** that they have witnessed any form of harrassment on social media with females with undergarduate beign the highest follow by Mbbsin view.

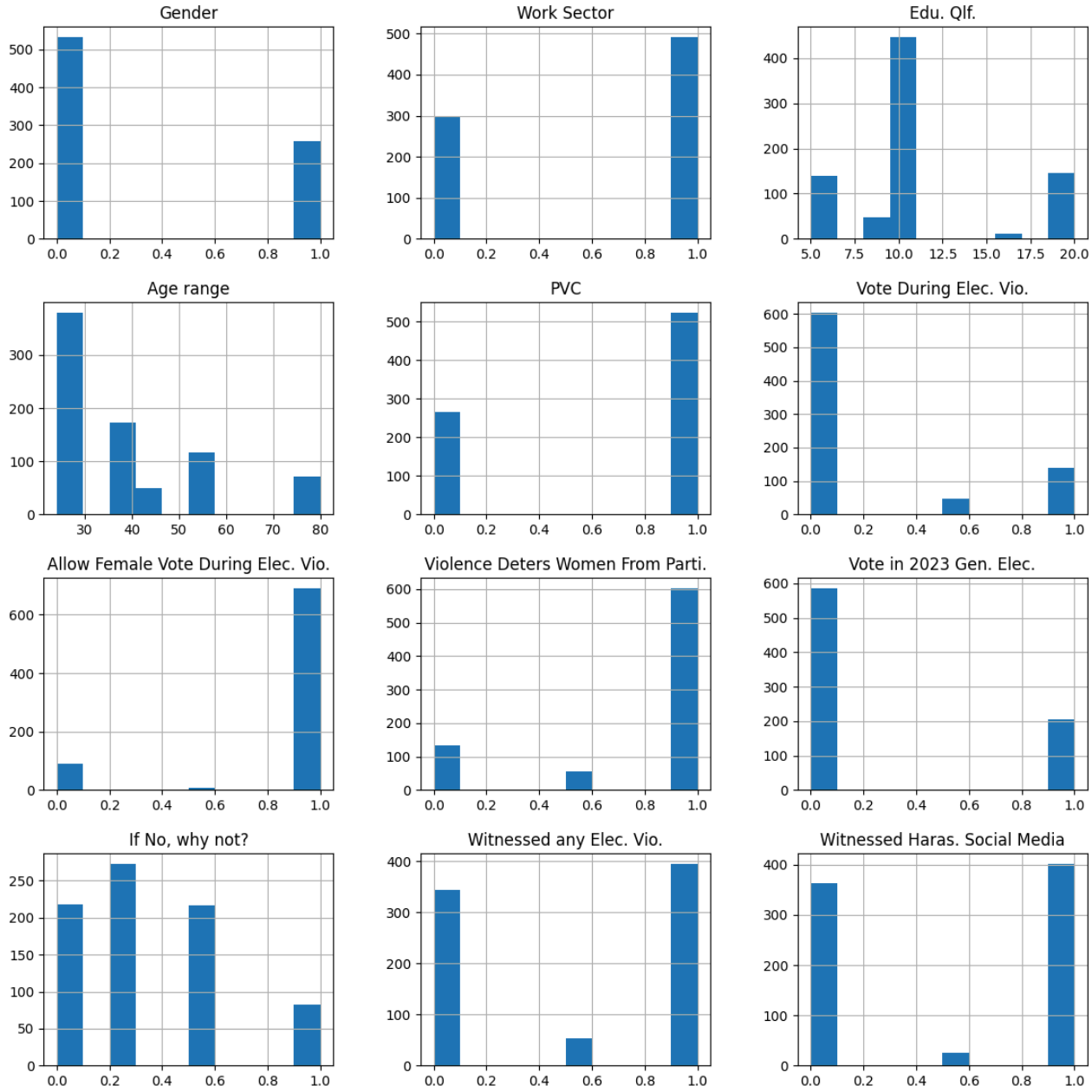
Summary of EDA

- The dataset has total of 533 Female and 258 Male respondent
- There are no missing values in the dataset
- High level of correlation exist between some of the variables.
- There are also outilers

Data Cleaning

- Timestamp column contains uniques ID for response. This column has been dropped.
- The target varaible is encoded to numeric.

In [151]:
```
# creating histograms
dataCorr.hist(figsize=(14, 14))
plt.show()
```



Let's find the percentage of outliers, in each column of the data, using IQR.

In [155]:
```
Q1 = dataCorr.quantile(0.25)  # To find the 25th percentile and 75th percentile.
Q3 = dataCorr.quantile(0.75)

IQR = Q3 - Q1  # Inter Quantile Range (75th perentile - 25th percentile)

lower = (
    Q1 - 1.5 * IQR
)  # Finding lower and upper bounds for all values. All values outside these bounds are outliers
upper = Q3 + 1.5 * IQR
```

In [158]:
```
lower
```

Out[158]:
```
Gender                              -1.50
Work Sector                         -1.50
Edu. Qlf.                           10.00
Age range                           -8.25
PVC                                 -1.50
Vote During Elec. Vio.               0.00
Allow Female Vote During Elec. Vio.  1.00
Violence Deters Women From Parti.    1.00
Vote in 2023 Gen. Elec.             -1.50
If No, why not?                     -0.75
Witnessed any Elec. Vio.            -1.50
Witnessed Haras. Social Media       -1.50
dtype: float64
```

In [159]:
```
upper
```

Out[159]:
```
Gender                               2.50
Work Sector                          2.50
Edu. Qlf.                           10.00
Age range                           77.75
PVC                                  2.50
Vote During Elec. Vio.               0.00
Allow Female Vote During Elec. Vio.  1.00
Violence Deters Women From Parti.    1.00
Vote in 2023 Gen. Elec.              2.50
If No, why not?                      1.25
Witnessed any Elec. Vio.             2.50
Witnessed Haras. Social Media        2.50
dtype: float64
```

In [156]:
```
(
    (dataCorr.select_dtypes(include=["float64", "int64"]) < lower)
    | (dataCorr.select_dtypes(include=["float64", "int64"]) > upper)
).sum() / len(data) * 100
```

Out[156]:
```
Gender                               0.000000
Work Sector                          0.000000
Edu. Qlf.                           44.627054
Age range                            8.975980
PVC                                  0.000000
Vote During Elec. Vio.              23.640961
Allow Female Vote During Elec. Vio. 12.642225
Violence Deters Women From Parti.    23.767383
Vote in 2023 Gen. Elec.              0.000000
If No, why not?                      0.000000
Witnessed any Elec. Vio.             0.000000
Witnessed Haras. Social Media        0.000000
dtype: float64
```

Observation

- After identifying outliers, we can decide whether to remove/treat them or not. It depends on one's approach, here we are not going to treat them as there will be outliers in real case scenario (in Education qualification, Vote During Elec. Vio., Allow Female Vote During Elec. Vio., etc) and we would want our model to learn the underlying pattern for such customers.

In [187]:
```
data1 = data.copy()
```

In [188]:
```
imputer = SimpleImputer(strategy="most_frequent")
```

In [189]:
```
X = data1.drop([ 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?'], axis=1)
y = data1[ 'Are you likely to prevent a "female" loved one from going to vote after violence occurs?']
```

In [190]:
```
# Splitting data into training, validation and test set:
# first we split data into 2 parts, say temporary and test

X_temp, X_test, y_temp, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1, stratify=y
)

# then we split the temporary set into train and validation

X_train, X_val, y_train, y_val = train_test_split(
    X_temp, y_temp, test_size=0.25, random_state=1, stratify=y_temp
)
print(X_train.shape, X_val.shape, X_test.shape)
```

```
(474, 12) (158, 12) (159, 12)
```

In [191]:
```
reqd_col_for_impute = ["Educational Qualification", "Gender", "Are you likely to vote when there is electoral violence a
```

In [192]:
```
# Fit and transform the train data
X_train[reqd_col_for_impute] = imputer.fit_transform(X_train[reqd_col_for_impute])

# Transform the validation data
X_val[reqd_col_for_impute] = imputer.transform(X_val[reqd_col_for_impute])

# Transform the test data
X_test[reqd_col_for_impute] = imputer.transform(X_test[reqd_col_for_impute])
```

In [193]:
```
# Checking that no column has missing values in train or test sets
print(X_train.isna().sum())
print("-" * 30)
print(X_val.isna().sum())
print("-" * 30)
print(X_test.isna().sum())
```

```
Gender
0
Work Sector
0
Educational Qualification
0
Age range
0
Do you have a permanent voters card?
0
Are you likely to vote when there is electoral violence around you?
0
Do you believe that violence deters women from participating in political activities such as rallies, elections, and campai
gns?    0
In your opinion, does violence impact the confidence of women in engaging in political activities?
0
Did you vote in 2023 General Elections?
0
If No, why not?
0
Have you ever witnessed any form of electoral violence during elections?
0
Have you ever witnessed any form of harassment on social media?
0
dtype: int64
------------------------------
Gender
0
Work Sector
0
Educational Qualification
0
Age range
0
Do you have a permanent voters card?
0
Are you likely to vote when there is electoral violence around you?
0
Do you believe that violence deters women from participating in political activities such as rallies, elections, and campai
gns?    0
In your opinion, does violence impact the confidence of women in engaging in political activities?
0
Did you vote in 2023 General Elections?
0
If No, why not?
0
Have you ever witnessed any form of electoral violence during elections?
0
Have you ever witnessed any form of harassment on social media?
0
dtype: int64
------------------------------
Gender
0
Work Sector
0
Educational Qualification
0
Age range
0
Do you have a permanent voters card?
0
Are you likely to vote when there is electoral violence around you?
0
Do you believe that violence deters women from participating in political activities such as rallies, elections, and campai
gns?    0
In your opinion, does violence impact the confidence of women in engaging in political activities?
0
Did you vote in 2023 General Elections?
0
If No, why not?
0
Have you ever witnessed any form of electoral violence during elections?
0
Have you ever witnessed any form of harassment on social media?
0
dtype: int64
```

- Here we confirm that there is no missing values / all missing values have been treated.

In [194]:
```
cols = X_train.select_dtypes(include=["object", "category"])
for i in cols.columns:
    print(X_train[i].value_counts())
    print("*" * 30)
```

```
Female    327
Male      147
Name: Gender, dtype: int64
******************************
Formal Sector (9-5 jobs, Professionals, Hybrid jobs)    294
Informal Sector (Artisans, Traders)                     180
Name: Work Sector, dtype: int64
******************************
HND, B.Sc.        275
Postgraduate       83
SSCE and below     78
ND, NCE            21
Mbbs in view        6
Undergraduate       6
Btech               5
Name: Educational Qualification, dtype: int64
******************************
18-30           223
31-40           101
51-60            69
60 and above     46
41-50            35
Name: Age range, dtype: int64
******************************
Yes    316
No     158
Name: Do you have a permanent voters card?, dtype: int64
******************************
No           356
Yes           84
Uncertain     34
Name: Are you likely to vote when there is electoral violence around you?, dtype: int64
******************************
Yes          349
No            87
Uncertain     38
Name: Do you believe that violence deters women from participating in political activities such as rallies, elections, and
campaigns?, dtype: int64
******************************
Yes          320
No           124
Uncertain     30
Name: In your opinion, does violence impact the confidence of women in engaging in political activities?, dtype: int64
******************************
No     345
Yes    129
Name: Did you vote in 2023 General Elections?, dtype: int64
******************************
Others                                                                  160
No PVC                                                                  130
Unavailable  (distance, health issues)                                   94
Electoral Violence                                                       49
Work (Journalist, Health officials, Security agents, Electoral officers)     41
Name: If No, why not?, dtype: int64
******************************
Yes          245
No           196
Uncertain     33
Name: Have you ever witnessed any form of electoral violence during elections?, dtype: int64
******************************
No           231
Yes          227
Uncertain     16
Name: Have you ever witnessed any form of harassment on social media?, dtype: int64
******************************
```

In [195]:
```
cols = X_val.select_dtypes(include=["object", "category"])
for i in cols.columns:
    print(X_val[i].value_counts())
    print("*" * 30)
```

```
Female    108
Male       50
Name: Gender, dtype: int64
******************************
Formal Sector (9-5 jobs, Professionals, Hybrid jobs)    97
Informal Sector (Artisans, Traders)                     61
Name: Work Sector, dtype: int64
******************************
HND, B.Sc.        80
SSCE and below    32
Postgraduate      31
ND, NCE            9
Mbbs in view       3
Btech              2
Undergraduate      1
Name: Educational Qualification, dtype: int64
******************************
18-30           78
31-40           35
51-60           22
60 and above    16
41-50            6
51-61            1
Name: Age range, dtype: int64
******************************
Yes    103
No      55
Name: Do you have a permanent voters card?, dtype: int64
******************************
No           119
Yes           29
Uncertain     10
Name: Are you likely to vote when there is electoral violence around you?, dtype: int64
******************************
Yes          128
No            23
Uncertain      7
Name: Do you believe that violence deters women from participating in political activities such as rallies, elections, and
campaigns?, dtype: int64
******************************
Yes          113
No            37
Uncertain      8
Name: In your opinion, does violence impact the confidence of women in engaging in political activities?, dtype: int64
******************************
No     121
Yes     37
Name: Did you vote in 2023 General Elections?, dtype: int64
******************************
Others                                                                  60
No PVC                                                                  48
Unavailable  (distance, health issues)                                  31
Electoral Violence                                                      13
Work (Journalist, Health officials, Security agents, Electoral officers)     6
Name: If No, why not?, dtype: int64
******************************
Yes          76
No           72
Uncertain    10
Name: Have you ever witnessed any form of electoral violence during elections?, dtype: int64
******************************
Yes          81
No           74
Uncertain     3
Name: Have you ever witnessed any form of harassment on social media?, dtype: int64
******************************
```

In [196]:
```
cols = X_test.select_dtypes(include=["object", "category"])
for i in cols.columns:
    print(X_train[i].value_counts())
    print("*" * 30)
```

```
Female    327
Male      147
Name: Gender, dtype: int64
******************************
Formal Sector (9-5 jobs, Professionals, Hybrid jobs)    294
Informal Sector (Artisans, Traders)                     180
Name: Work Sector, dtype: int64
******************************
HND, B.Sc.        275
Postgraduate       83
SSCE and below     78
ND, NCE            21
Mbbs in view        6
Undergraduate       6
Btech               5
Name: Educational Qualification, dtype: int64
******************************
18-30           223
31-40           101
51-60            69
60 and above     46
41-50            35
Name: Age range, dtype: int64
******************************
Yes    316
No     158
Name: Do you have a permanent voters card?, dtype: int64
******************************
No           356
Yes           84
Uncertain     34
Name: Are you likely to vote when there is electoral violence around you?, dtype: int64
******************************
Yes          349
No            87
Uncertain     38
Name: Do you believe that violence deters women from participating in political activities such as rallies, elections, and
campaigns?, dtype: int64
******************************
Yes          320
No           124
Uncertain     30
Name: In your opinion, does violence impact the confidence of women in engaging in political activities?, dtype: int64
******************************
No     345
Yes    129
Name: Did you vote in 2023 General Elections?, dtype: int64
******************************
Others                                                                  160
No PVC                                                                  130
Unavailable  (distance, health issues)                                   94
Electoral Violence                                                       49
Work (Journalist, Health officials, Security agents, Electoral officers)     41
Name: If No, why not?, dtype: int64
******************************
Yes          245
No           196
Uncertain     33
Name: Have you ever witnessed any form of electoral violence during elections?, dtype: int64
******************************
No           231
Yes          227
Uncertain     16
Name: Have you ever witnessed any form of harassment on social media?, dtype: int64
******************************
```

Encoding categorical variables

In [197]:
```
X_train = pd.get_dummies(X_train, drop_first=True)
X_val = pd.get_dummies(X_val, drop_first=True)
X_test = pd.get_dummies(X_test, drop_first=True)
print(X_train.shape, X_val.shape, X_test.shape)
```

```
(474, 28) (158, 29) (159, 28)
```

- After encoding there are 28 columns.

In [198]:
```
X_train.head()
```

Out[198]:

	Gender_Male	Work Sector_Informal Sector (Artisans, Traders)	Educational Qualification_HND, B.Sc.	Educational Qualification_Mbbs in view	Educational Qualification_ND, NCE	Educational Qualification_Postgraduate	Educational Qualification_SSCE and below	Qualification_
570	0	0	0	0	0	0	1	
745	1	1	1	0	0	0	0	
238	0	1	1	0	0	0	0	
106	0	1	1	0	0	0	0	
257	1	1	0	0	0	1	0	

Model Building

Model evaluation criterion

The nature of predictions made by the classification model will translate as follows:

- True positives (TP) are failures correctly predicted by the model.
- False negatives (FN) are real failures in a generator where there is no detection by model.
- False positives (FP) are failure detections in a generator where there is no failure.
- Accuracy
- Precision
- Recall
- F1-Score

Which metric to optimize?

- We need to choose the metric which will ensure that the maximum number of generator failures are predicted correctly by the model.
- We would want Recall to be maximized as greater the Recall, the higher the chances of minimizing false negatives.
- We want to minimize false negatives because if a model predicts that a machine will have no failure when there will be a failure, it will increase the maintenance cost.

Let's define a function to output different metrics (including recall) on the train and test set and a function to show confusion matrix so that we do not have to use the same code repetitively while evaluating models.

In [199]:
```
# defining a function to compute different metrics to check performance of a classification model built using sklearn
def model_performance_classification_sklearn(model, predictors, target):
    """
    Function to compute different metrics to check classification model performance

    model: classifier
    predictors: independent variables
    target: dependent variable
    """

    # predicting using the independent variables
    pred = model.predict(predictors)

    acc = accuracy_score(target, pred)  # to compute Accuracy
    recall = recall_score(target, pred)  # to compute Recall
    precision = precision_score(target, pred)  # to compute Precision
    f1 = f1_score(target, pred)  # to compute F1-score

    # creating a dataframe of metrics
    df_perf = pd.DataFrame(
        {"Accuracy": acc, "Recall": recall, "Precision": precision, "F1": f1,},
        index=[0],
    )

    return df_perf
```

In [200]:
```
def confusion_matrix_sklearn(model, predictors, target):
    """
    To plot the confusion_matrix with percentages

    model: classifier
    predictors: independent variables
    target: dependent variable
    """
    y_pred = model.predict(predictors)
    cm = confusion_matrix(target, y_pred)
    labels = np.asarray(
        [
            ["{0:0.0f}".format(item) + "\n{0:.2%}".format(item / cm.flatten().sum())]
            for item in cm.flatten()
        ]
    ).reshape(2, 2)

    plt.figure(figsize=(6, 4))
    sns.heatmap(cm, annot=labels, fmt="")
    plt.ylabel("True label")
    plt.xlabel("Predicted label")
```

Model with original data

In [202]:
```
X_train
```

Out[202]:

	Gender_Male	Work Sector_Informal Sector (Artisans, Traders)	Educational Qualification_HND, B.Sc.	Educational Qualification_Mbbs in view	Educational Qualification_ND, NCE	Educational Qualification_Postgraduate	Educational Qualification_SSCE and below	Qualification_
570	0	0	0	0	0	0	1	
745	1	1	1	0	0	0	0	
238	0	1	1	0	0	0	0	
106	0	1	1	0	0	0	0	
257	1	1	0	0	0	1	0	
...	...	...	...	...	...	...	...	
638	0	0	1	0	0	0	0	
555	1	1	1	0	0	0	0	
207	0	0	1	0	0	0	0	
444	0	1	1	0	0	0	0	
720	1	0	0	0	0	0	1	

474 rows × 28 columns

In [205]:
```
X_val
```

Out[205]:

	Gender_Male	Work Sector_Informal Sector (Artisans, Traders)	Educational Qualification_HND, B.Sc.	Educational Qualification_Mbbs in view	Educational Qualification_ND, NCE	Educational Qualification_Postgraduate	Educational Qualification_SSCE and below	Qualificatio
722	0	0	0	0	0	1	0	
9	0	0	1	0	0	0	0	
450	1	0	0	0	0	1	0	
136	0	0	0	0	0	1	0	

In [207]:
```
X_val.drop([ 'Age range_51-61'], axis=1,inplace=True)
X_val
```

Out[207]:

	Gender_Male	Work Sector_Informal Sector (Artisans, Traders)	Educational Qualification_HND, B.Sc.	Educational Qualification_Mbbs in view	Educational Qualification_ND, NCE	Educational Qualification_Postgraduate	Educational Qualification_SSCE and below	Qualificatic
722	0	0	0	0	0	1	0	
9	0	0	1	0	0	0	0	
450	1	0	0	0	0	1	0	
136	0	0	0	0	0	1	0	

In [228]:
```
models = []  # Empty list to store all the models

# Appending models into the list
models.append(("Logistic regression", LogisticRegression(random_state=1)))


print("\n" "Training Performance:" "\n")
for name, model in models:
    model.fit(X_train, y_train)
    scores = recall_score(y_train, model.predict(X_train), average='weighted')
    print("{}: {}".format(name, scores))

print("\n" "Validation Performance:" "\n")

for name, model in models:
    model.fit(X_train, y_train)
    scores_val = recall_score(y_val, model.predict(X_val), average='weighted')
    print("{}: {}".format(name, scores_val))
```

```
Training Performance:

Logistic regression: 0.9261603375527426

Validation Performance:

Logistic regression: 0.9430379746835443
```

Performance Analysis

- Logistic Regression for predicting participation of Female in Election process

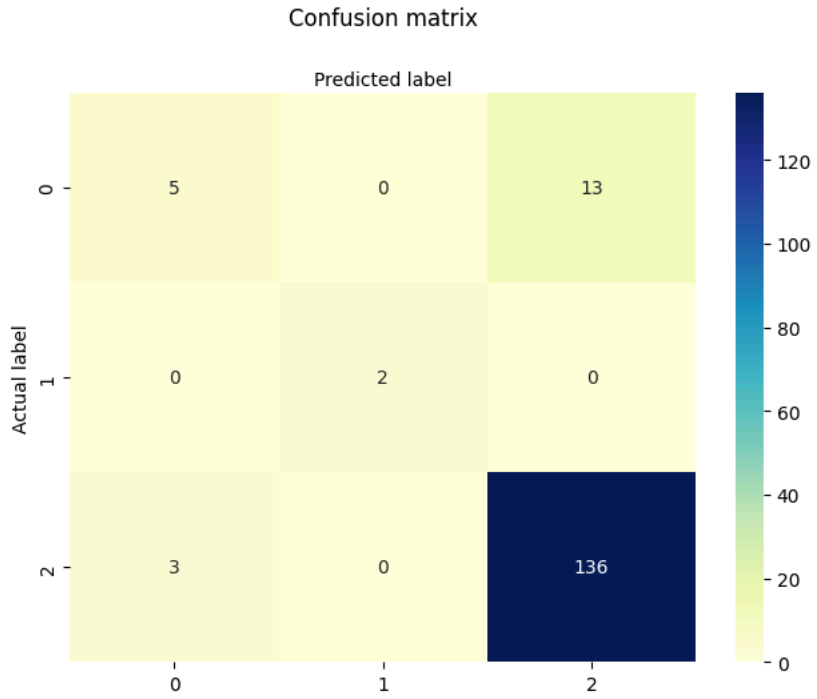
In [219]:
```
# instantiate the model (using the default parameters)
# LogisticRegression(random_state=1)
logreg = LogisticRegression(random_state=16)

# fit the model with data
logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_test)
```

In [220]:
```
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

Out[220]:
```
array([[  5,   0,  13],
       [  0,   2,   0],
       [  3,   0, 136]], dtype=int64)
```

In [222]:
```
# import required modules
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')

# Text(0.5,257.44,'Predicted Label');
```

Out[222]: Text(0.5, 427.9555555555555, 'Predicted label')



In [224]:
```
from sklearn.metrics import classification_report
target_names = ['Yes', 'No', 'Uncertain']
print(classification_report(y_test, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

         Yes       0.62      0.28      0.38        18
          No       1.00      1.00      1.00         2
   Uncertain       0.91      0.98      0.94       139

    accuracy                           0.90       159
   macro avg       0.85      0.75      0.78       159
weighted avg       0.88      0.90      0.88       159
```

DESCRIPTION OF THE WORK SO FAR

- **STEP 1:** EXPLORATORY DATA ANALYSIS (EDA)
  - the EDA analysis provide proper understanding of the data
  - show existence of outliers
- **STEP 2:** SPLITING OF DATA
  - Split data into Train and Test Set
- **STEP 3:** APPLY MINMAX
  - Apply MinMax Scaler on Train Set
  - Apply MinMax Scaler on Test Set
- **STEP 4:** TRAIN MODEL
  - Train the Logictic Regression Model on the Train Set
- **STEP 5:** EVALUATE
  - Evaluate the Model performance on the Test Set
  - metric include Accuracy, Recall, F1-score, Precision
  - the model recorded the following performance
    1. **Accuracy of 90%**,
    2. **F1-score of 88%**,
    3. **Recall of 90%** and

4. **Precision 88%**