

Data Analysis and Visualisation Project

Title: Tobacco Consumption and analysis

Project Leader:

Tanisha Kriplani, 3422

Team Members:

- Diksha, 3443
- Muskan Gupta, 3458
- Sourabh Takshak, 3486
- Vedanti Kiran, 3502
- And Munkhzul, 3500

Objective/Aim:

To analyse the consumption of different types of tobacco in various states of India.

Technical Details:

Python language with libraries numpy, pandas, matplotlib, seaborn.

Innovativeness and Usefulness:

To get a vivid picture of different types of tobacco consumption in India, so they can have a cutting-edge perspective on public health trends enabling stakeholders to formulate informed policies and strategies. This examination is pivotal for understanding the behavioural patterns of consumers, which can lead to innovative interventions that curb its use and mitigate associated health risks.

Market potential and Competitive Advantage:

Examining the data on tobacco usage in India not only uncovers significant market opportunities in one of the globe's most densely populated nations, but also furnishes critical insights, granting stakeholders a distinct competitive edge to refine their approaches and respond to shifting consumer trends adeptly.

Introduction

India's diverse and expansive population has positioned it as one of the world's most significant markets for tobacco products. With consumption patterns ranging from traditional cigarettes and bidis to smokeless forms like gutka, the landscape is multifaceted. As health concerns related to tobacco use escalate, it becomes imperative to understand its consumption trends in depth. This Data Analysis and Visualization project aims to uncover these patterns, providing a detailed perspective on regional variations, age-related preferences, and other key factors. Our insights will serve as valuable information for policymakers, healthcare professionals, and those aiming to combat the adverse effects of tobacco in India.

This project uses the dataset imported from Kaggle. The dataset is available at <https://www.kaggle.com/datasets/anshtanwar/youth-tobacco-survey>. This dataset includes a csv file, representing the consumption of various types of tobacco in different states of India. In order to draw useful insights, this project analyses the different types of tobacco consumption in India.

Current Status of Development: Completed.

1. Importing Libraries

```
In [1]: import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns
```

Reading dataset CSV file

```
In [2]: data = pd.read_csv("datafile.csv")
```

2. Exploring the dataset

Checking the rows and columns of the data

```
In [3]: #Display first five rows  
data.head()
```

Out[3]:

	State/UT	Area	Ever tobacco users (%)	Current tobacco users (%)	Ever tobacco smokers (%)	Current tobacco smokers (%)	Ever cigarette users (%)	Current cigarette users (%)	Ever bidi users (%)
0	India	Total	18.1	8.5	11.0	7.3	4.6	2.6	4.1
1	India	Urban	13.5	5.5	8.1	4.5	4.1	2.3	3.7
2	India	Rural	19.5	9.4	11.8	8.1	4.7	2.6	4.3
3	Andaman and Nicobar Islands	Total	21.5	4.4	8.8	3.7	4.4	1.3	2.7
4	Andaman and Nicobar Islands	Urban	13.3	3.4	6.5	2.6	3.7	1.3	2.0

5 rows × 59 columns

In [4]: *#Dimensions of dataset*
data.shape

Out[4]: (107, 59)

In [5]: *#information of dataset*
data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 59 columns):
 #   Column          Non-Null Count  Dtype  
--- 
 0   State/UT        107 non-null    object 
 1   Area            107 non-null    object 
 2   Ever tobacco users (%) 107 non-null    float64
 3   Current tobacco users (%) 107 non-null    float64
 4   Ever tobacco smokers (%) 107 non-null    float64
 5   Current tobacco smokers (%) 107 non-null    float64
 6   Ever cigarette users (%) 107 non-null    float64
 7   Current cigarette users (%) 107 non-null    float64
 8   Ever bidi users (%) 107 non-null    float64
 9   Current bidi users (%) 107 non-null    float64
 10  Ever smokeless tobacco users (%) 107 non-null    float64
 11  Current smokeless tobacco users (%) 107 non-null    float64
 12  Ever users of paan masala together with tobacco (%) 107 non-null    float64
 13  Never cigarette smokers susceptible to cigarette use in future (%) 107 non-null    float64
 14  Median age of initiation of Cigarette (in years) 104 non-null    float64
 15  Median age of initiation of Bidi (in years) 106 non-null    float64
 16  Median age of initiation of smokeless tobacco (in years) 94 non-null    float64
 17  Awareness about e-cigarette (%) 107 non-null    float64
 18  Ever e-cigarette use (%) 107 non-null    float64
 19  Ever tobacco smokers who quit in last 12 months (%) 107 non-null    float64
 20  Current tobacco smokers who tried to quit smoking in the past 12 mon 107 no
ths (%) 
 21  Current tobacco smokers who wanted to quit smoking now (%) 107 non-null    float64
 22  Ever smokeless tobacco users who quit in last 12 months (%) 107 non-null    float64
 23  Current smokeless tobacco users who tried to quit tobacco in the pas 107 no
t 12 months (%) 
 24  Current smokeless tobacco users who wanted to quit tobacco now
```

(%) 107 no
n-null float64
25 Exposure to tobacco smoke at home/public place (%)
107 non-null float64
26 Exposure to tobacco smoke at home (%)
107 non-null float64
27 Exposure to tobacco smoke inside any enclosed public places (%)
107 non-null float64
28 Exposure to tobacco smoke at any outdoor public places (%)
107 non-null float64
29 Students who saw anyone smoking inside the school building or outside school property (%) 107 no
n-null float64
30 Major source of tobacco product- Cigarette: Store (%)
107 non-null float64
31 Major source of tobacco product- Cigarette: Paan shop (%)
107 non-null float64
32 Major source of tobacco product- Bidi: Store (%)
107 non-null float64
33 Major source of tobacco product- Bidi: Paan shop (%)
107 non-null float64
34 Major source of tobacco product- Smokeless tobacco: Store (%)
107 non-null float64
35 Major source of tobacco product- Smokeless tobacco: Paan shop (%)
107 non-null float64
36 Current cigarette smokers who bought cigarettes from a store, paan shop, street vendor, or vending machine (%) 107 no
n-null float64
37 Current bidi smokers who bought bidi from a store, paan shop, or street vendor (%) 107 no
n-null float64
38 Refused sale of cigarette because of age in past 30 days (%)
107 non-null float64
39 Refused sale of bidi because of age in past 30 days (%)
107 non-null float64
40 Refused sale of smokeless tobacco because of age in past 30 days (%) 107 no
n-null float64
41 Bought Cigarette as individual sticks in past 30 days (%)
107 non-null object
42 Bought Bidi as individual sticks in past 30 days (%)
107 non-null float64
43 Students who noticed anti-tobacco messages anywhere in past 30 days (%) 107 non-null float64
44 Students who noticed anti-tobacco messages in mass media in past 30 days (%) 107 no
n-null float64
45 Students who noticed anti-tobacco messages at sporting, fairs, concerts, community events or social gatherings in past 30 days (%) 107 no
n-null float64
46 Students who noticed health warnings on any tobacco product/cigarette packages in past 30 days (%) 107 no
n-null float64
47 Students who saw tobacco advertisements anywhere in past 30 days (%) 107
non-null float64
48 Students who saw anyone using tobacco on mass media in past 30 days

```
(%) 107 non
- null object
49 Students who noticed cigarette advertisements/promotions at point of
sale in past 30 days (%) 107 non
- null float64
50 Students who were taught in class about harmful effects of tobacco u
se during past 12 months (%) 107 no
n-null float64
51 Students who thought it is difficult to quit once someone starts smo
king tobacco (%) 107 no
n-null float64
52 Students who thought other people's tobacco smoking is harmful to th
em (%) 107 no
n-null float64
53 Students who favoured ban on smoking inside enclosed public places
(%) 107 n
on-null float64
54 Students who favoured ban on smoking at outdoor public places (%)
107 non-null float64
55 School heads aware of COTPA, 2003 (%) 107 no
107 non-null float64
56 Schools authorized by the state government to collect fine for viola
tion under Section-6 of the COTPA, 2003 (%) 107 no
n-null float64
57 Schools followed 'tobacco-free school' guidelines (%) 107 no
107 non-null float64
58 Schools aware of the policy for displaying 'tobacco-free school' bo
ard (%) 107 no
n-null float64
dtypes: float64(55), object(4)
memory usage: 49.4+ KB
```

```
In [6]: #Check columns of dataset
data.columns
```

```
Out[6]: Index(['State/UT', 'Area', 'Ever tobacco users (%)',
   'Current tobacco users (%)', 'Ever tobacco smokers (%)',
   'Current tobacco smokers (%)', 'Ever cigarette users (%)',
   'Current cigarette users (%)', 'Ever bidi users (%)',
   'Current bidi users (%)', 'Ever smokeless tobacco users (%)',
   'Current smokeless tobacco users (%)',
   'Ever users of paan masala together with tobacco (%)',
   'Never cigarette smokers susceptible to cigarette use in future
   (%)',
   'Median age of initiation of Cigarette (in years)',
   'Median age of initiation of Bidi (in years)',
   'Median age of initiation of smokeless tobacco (in years)',
   'Awareness about e-cigarette (%)', 'Ever e-cigarette use (%)',
   'Ever tobacco smokers who quit in last 12 months (%)',
   'Current tobacco smokers who tried to quit smoking in the past 12
months (%)',
   'Current tobacco smokers who wanted to quit smoking now (%)',
   'Ever smokeless tobacco users who quit in last 12 months (%)',
   'Current smokeless tobacco users who tried to quit tobacco in the
past 12 months (%)',
   'Current smokeless tobacco users who wanted to quit tobacco now
   (%)',
   'Exposure to tobacco smoke at home/public place (%)',
   'Exposure to tobacco smoke at home (%)',
   'Exposure to tobacco smoke inside any enclosed public places
   (%)',
   'Exposure to tobacco smoke at any outdoor public places (%)',
   'Students who saw anyone smoking inside the school building or ou
tside school property (%)',
   'Major source of tobacco product- Cigarette: Store (%)',
   'Major source of tobacco product- Cigarette: Paan shop (%)',
   'Major source of tobacco product- Bidi: Store (%)',
   'Major source of tobacco product- Bidi: Paan shop (%)',
   'Major source of tobacco product- Smokeless tobacco: Store (%)',
   'Major source of tobacco product- Smokeless tobacco: Paan shop
   (%)',
   'Current cigarette smokers who bought cigarettes from a store, paa
n shop, street vendor, or vending machine (%)',
   'Current bidi smokers who bought bidi from a store, paan shop, or
street vendor (%)',
   'Refused sale of cigarette because of age in past 30 days (%)',
   'Refused sale of bidi because of age in past 30 days (%)',
   'Refused sale of smokeless tobacco because of age in past 30 days
   (%)',
   'Bought Cigarette as individual sticks in past 30 days (%)',
   'Bought Bidi as individual sticks in past 30 days (%)',
   'Students who noticed anti-tobacco messages anywhere in past 30 da
ys (%)',
   'Students who noticed anti-tobacco messages in mass media in past
30 days (%)',
   'Students who noticed anti-tobacco messages at sporting, fairs, c
oncerts, community events or social gatherings in past 30 days (%)',
   'Students who noticed health warnings on any tobacco product/cigar
ette packages in past 30 days (%)',
   'Students who saw tobacco advertisements anywhere in past 30 days
   (%)',
   'Students who saw anyone using tobacco on mass media in past 30 da
ys (%)',
```

```

'Students who noticed cigarette advertisements/promotions at point
of sale in past 30 days (%)',
'Students who were taught in class about harmful effects of tobacco
use during past 12 months (%)',
'Students who thought it is difficult to quit once someone starts
smoking tobacco (%)',
'Students who thought other people's tobacco smoking is harmful to
them (%)',
'Students who favoured ban on smoking inside enclosed public places
(%)',
'Students who favoured ban on smoking at outdoor public places
(%)',
'School heads aware of COTPA, 2003 (%)',
'Schools authorized by the state government to collect fine for vi
olation under Section-6 of the COTPA, 2003 (%)',
'Schools followed 'tobacco-free school' guidelines (%)',
'Schools aware of the policy for displaying 'tobacco-free school'
board (%)'],
dtype='object')

```

In [7]: `data.describe()`

Out[7]:

	Ever tobacco users (%)	Current tobacco users (%)	Ever tobacco smokers (%)	Current tobacco smokers (%)	Ever cigarette users (%)	Current cigarette users (%)	Ever t users (%)
count	107.000000	107.000000	107.000000	107.000000	107.000000	107.000000	107.000000
mean	22.766355	11.361682	14.852336	9.095327	9.910280	5.876636	6.9891
std	19.375305	14.912852	17.127818	12.168366	15.507466	9.949804	9.5294
min	0.100000	0.000000	0.100000	0.000000	0.000000	0.000000	0.0000
25%	10.700000	3.250000	5.700000	2.500000	2.400000	1.050000	2.3500
50%	15.300000	5.500000	8.800000	4.300000	4.200000	2.300000	3.5000
75%	26.300000	8.550000	12.450000	7.600000	7.150000	4.350000	5.5000
max	91.100000	69.200000	84.500000	59.600000	77.100000	47.800000	50.2000

8 rows × 55 columns

3.Data Cleaning

3.1 Dropping irrelevant columns

```
In [8]: data_1 = data.drop(data.columns[29:], axis=1)
```

```
In [9]: #After drop no. of columns  
data_1.columns
```

```
Out[9]: Index(['State/UT', 'Area', 'Ever tobacco users (%)',  
               'Current tobacco users (%)', 'Ever tobacco smokers (%)',  
               'Current tobacco smokers (%)', 'Ever cigarette users (%)',  
               'Current cigarette users (%)', 'Ever bidi users (%)',  
               'Current bidi users (%)', 'Ever smokeless tobacco users (%)',  
               'Current smokeless tobacco users (%)',  
               'Ever users of paan masala together with tobacco (%)',  
               'Never cigarette smokers susceptible to cigarette use in future  
(%)',  
               'Median age of initiation of Cigarette (in years)',  
               'Median age of initiation of Bidi (in years)',  
               'Median age of initiation of smokeless tobacco (in years)',  
               'Awareness about e-cigarette (%)', 'Ever e-cigarette use (%)',  
               'Ever tobacco smokers who quit in last 12 months (%)',  
               'Current tobacco smokers who tried to quit smoking in the past 12  
months (%)',  
               'Current tobacco smokers who wanted to quit smoking now (%)',  
               'Ever smokeless tobacco users who quit in last 12 months (%)',  
               'Current smokeless tobacco users who tried to quit tobacco in the  
past 12 months (%)',  
               'Current smokeless tobacco users who wanted to quit tobacco now  
(%)',  
               'Exposure to tobacco smoke at home/public place (%)',  
               'Exposure to tobacco smoke at home (%)',  
               'Exposure to tobacco smoke inside any enclosed public places  
(%)',  
               'Exposure to tobacco smoke at any outdoor public places (%)'],  
               dtype='object')
```

3.2 Handling Missing Values

Checking for the null values in each column

```
In [10]: data_1.isnull().sum()
```

Out[10]: State/UT
0
Area
0
Ever tobacco users (%)
0
Current tobacco users (%)
0
Ever tobacco smokers (%)
0
Current tobacco smokers (%)
0
Ever cigarette users (%)
0
Current cigarette users (%)
0
Ever bidi users (%)
0
Current bidi users (%)
0
Ever smokeless tobacco users (%)
0
Current smokeless tobacco users (%)
0
Ever users of paan masala together with tobacco (%)
0
Never cigarette smokers susceptible to cigarette use in future (%)
0
Median age of initiation of Cigarette (in years)
3
Median age of initiation of Bidi (in years)
1
Median age of initiation of smokeless tobacco (in years)
13
Awareness about e-cigarette (%)
0
Ever e-cigarette use (%)
0
Ever tobacco smokers who quit in last 12 months (%)
0
Current tobacco smokers who tried to quit smoking in the past 12 months (%) 0
Current tobacco smokers who wanted to quit smoking now (%)
0
Ever smokeless tobacco users who quit in last 12 months (%)
0
Current smokeless tobacco users who tried to quit tobacco in the past 12 months (%) 0
Current smokeless tobacco users who wanted to quit tobacco now (%)
0
Exposure to tobacco smoke at home/public place (%)
0
Exposure to tobacco smoke at home (%)
0
Exposure to tobacco smoke inside any enclosed public places (%)
0
Exposure to tobacco smoke at any outdoor public places (%)

```
0  
dtype: int64
```

```
In [11]: data_1.isnull().sum().sum()
```

```
Out[11]: 17
```

```
In [ ]: #Percentage of Null values  
for col in data_1.columns:  
    null = data_1[col].isnull().sum()  
    percentage = (null/len(data_1))*100  
    print(col, " null percentage : ", round(percentage), "%")
```

```
State/UT null percentage : 0 %  
Area null percentage : 0 %  
Ever tobacco users (%) null percentage : 0 %  
Current tobacco users (%) null percentage : 0 %  
Ever tobacco smokers (%) null percentage : 0 %  
Current tobacco smokers (%) null percentage : 0 %  
Ever cigarette users (%) null percentage : 0 %  
Current cigarette users (%) null percentage : 0 %  
Ever bidi users (%) null percentage : 0 %  
Current bidi users (%) null percentage : 0 %  
Ever smokeless tobacco users (%) null percentage : 0 %  
Current smokeless tobacco users (%) null percentage : 0 %  
Ever users of paan masala together with tobacco (%) null percentage : 0 %  
Never cigarette smokers susceptible to cigarette use in future (%) null percentage : 0 %  
Median age of initiation of Cigarette (in years) null percentage : 3 %  
Median age of initiation of Bidi (in years) null percentage : 1 %  
Median age of initiation of smokeless tobacco (in years) null percentage : 12 %  
Awareness about e-cigarette (%) null percentage : 0 %  
Ever e-cigarette use (%) null percentage : 0 %  
Ever tobacco smokers who quit in last 12 months (%) null percentage : 0 %  
Current tobacco smokers who tried to quit smoking in the past 12 months (%) null percentage : 0 %  
Current tobacco smokers who wanted to quit smoking now (%) null percentage : 0 %  
Ever smokeless tobacco users who quit in last 12 months (%) null percentage : 0 %  
Current smokeless tobacco users who tried to quit tobacco in the past 12 months (%) null percentage : 0 %  
Current smokeless tobacco users who wanted to quit tobacco now (%) null percentage : 0 %  
Exposure to tobacco smoke at home/public place (%) null percentage : 0 %  
Exposure to tobacco smoke at home (%) null percentage : 0 %  
Exposure to tobacco smoke inside any enclosed public places (%) null percentage : 0 %  
Exposure to tobacco smoke at any outdoor public places (%) null percentage : 0 %
```

Fill null values

```
In [12]: b = data_1["Median age of initiation of Cigarette (in years)"].mean()
```

```
In [13]: data_1["Median age of initiation of Cigarette (in years)"].fillna(b,inplace=True)

In [14]: c = data_1["Median age of initiation of Bidi (in years)"].mean()

In [15]: data_1["Median age of initiation of Bidi (in years)"].fillna(c, inplace=True)

In [16]: d = data_1["Median age of initiation of smokeless tobacco (in years)"].mean()

In [17]: data_1["Median age of initiation of smokeless tobacco (in years)"].fillna(d,inplace=True)
```

Checking Null Values Now

```
In [18]: data_1.isnull().sum()
```

Out[18]: State/UT
0
Area
0
Ever tobacco users (%)
0
Current tobacco users (%)
0
Ever tobacco smokers (%)
0
Current tobacco smokers (%)
0
Ever cigarette users (%)
0
Current cigarette users (%)
0
Ever bidi users (%)
0
Current bidi users (%)
0
Ever smokeless tobacco users (%)
0
Current smokeless tobacco users (%)
0
Ever users of paan masala together with tobacco (%)
0
Never cigarette smokers susceptible to cigarette use in future (%)
0
Median age of initiation of Cigarette (in years)
0
Median age of initiation of Bidi (in years)
0
Median age of initiation of smokeless tobacco (in years)
0
Awareness about e-cigarette (%)
0
Ever e-cigarette use (%)
0
Ever tobacco smokers who quit in last 12 months (%)
0
Current tobacco smokers who tried to quit smoking in the past 12 months (%) 0
Current tobacco smokers who wanted to quit smoking now (%)
0
Ever smokeless tobacco users who quit in last 12 months (%)
0
Current smokeless tobacco users who tried to quit tobacco in the past 12 months (%) 0
Current smokeless tobacco users who wanted to quit tobacco now (%)
0
Exposure to tobacco smoke at home/public place (%)
0
Exposure to tobacco smoke at home (%)
0
Exposure to tobacco smoke inside any enclosed public places (%)
0
Exposure to tobacco smoke at any outdoor public places (%)

```
0  
dtype: int64
```

```
In [19]: data_1.isnull().sum().sum()
```

```
Out[19]: 0
```

```
In [20]: data_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 29 columns):
 #   Column          Non-Null Count  Dtype  
--- 
 0   State/UT        107 non-null    object 
 1   Area            107 non-null    object 
 2   Ever tobacco users (%) 107 non-null    float64
 3   Current tobacco users (%) 107 non-null    float64
 4   Ever tobacco smokers (%) 107 non-null    float64
 5   Current tobacco smokers (%) 107 non-null    float64
 6   Ever cigarette users (%) 107 non-null    float64
 7   Current cigarette users (%) 107 non-null    float64
 8   Ever bidi users (%) 107 non-null    float64
 9   Current bidi users (%) 107 non-null    float64
 10  Ever smokeless tobacco users (%) 107 non-null    float64
 11  Current smokeless tobacco users (%) 107 non-null    float64
 12  Ever users of paan masala together with tobacco (%) 107 non-null    float64
 13  Never cigarette smokers susceptible to cigarette use in future (%) 107 non-null    float64
 14  Median age of initiation of Cigarette (in years) 107 non-null    float64
 15  Median age of initiation of Bidi (in years) 107 non-null    float64
 16  Median age of initiation of smokeless tobacco (in years) 107 non-null    float64
 17  Awareness about e-cigarette (%) 107 non-null    float64
 18  Ever e-cigarette use (%) 107 non-null    float64
 19  Ever tobacco smokers who quit in last 12 months (%) 107 non-null    float64
 20  Current tobacco smokers who tried to quit smoking in the past 12 months (%) 107 non-null    float64
 21  Current tobacco smokers who wanted to quit smoking now (%) 107 non-null    float64
 22  Ever smokeless tobacco users who quit in last 12 months (%) 107 non-null    float64
 23  Current smokeless tobacco users who tried to quit tobacco in the past 12 months (%) 107 non-null    float64
 24  Current smokeless tobacco users who wanted to quit tobacco now (%) 107 non-null    float64
 25  Exposure to tobacco smoke at home/public place (%)
```

```
107 non-null    float64
26  Exposure to tobacco smoke at home      (%)
107 non-null    float64
27  Exposure to tobacco smoke inside any enclosed public places  (%)
107 non-null    float64
28  Exposure to tobacco smoke at any outdoor public places  (%)
107 non-null    float64
dtypes: float64(27), object(2)
memory usage: 24.4+ KB
```

3.3 Handling Duplicate Values

Checking for duplicate values

```
In [21]: data_1.duplicated().sum()
```



```
Out[21]: 0
```

```
In [23]: data_2 = data_1.copy()
```

3.4 Handling Outliers

Checking outliers through Boxplot

```
In [ ]: data_1.to_csv('data_22.csv')
```

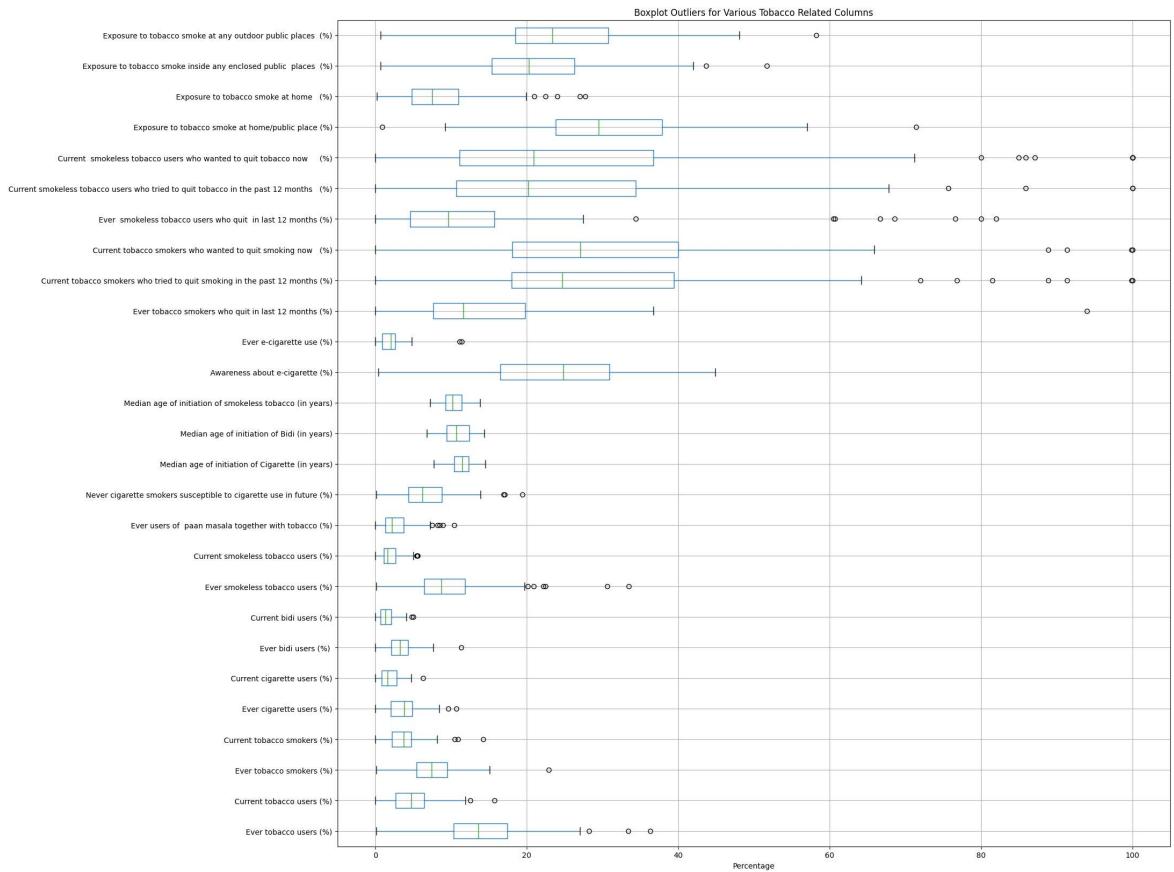


```
In [24]: # Calculate the upper and lower limits
Q1 = data_2['Current tobacco users (%)'].quantile(0.25)
Q3 = data_2['Current tobacco users (%)'].quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5*IQR
upper = Q3 + 1.5*IQR

# Create arrays of Boolean values indicating the outlier rows
upper_array = np.where(data_2['Current tobacco users (%)']>=upper)[0]
lower_array = np.where(data_2['Current tobacco users (%)']<=lower)[0]

# Removing the outliers
data_2.drop(index=upper_array, inplace=True)
data_2.drop(index=lower_array, inplace=True)
```

```
In [25]: data_2.boxplot(figsize=(20, 20), vert=False)
plt.title('Boxplot Outliers for Various Tobacco Related Columns')
plt.xlabel('Percentage')
plt.show()
```



```
In [26]: data_2.shape
```

```
Out[26]: (85, 29)
```

Query 1

Find states where over 50% of the population have never used tobacco.

```
In [ ]: dr = data_2.rename(columns= {'Ever tobacco users (%)':'Ever_tobacco_users'})
non_users = dr[dr['Ever_tobacco_users'] < 50]
print(non_users[["State/UT", "Ever_tobacco_users"]])
```

	State/UT	Ever_tobacco_users
0	India	18.1
1	India	13.5
2	India	19.5
3	Andaman and Nicobar Islands	21.5
4	Andaman and Nicobar Islands	13.3
..
98	Tripura	23.8
100	Uttar Pradesh	6.6
104	Uttarakhand	28.2
105	West Bengal	17.3
106	West Bengal	18.0

[85 rows x 2 columns]

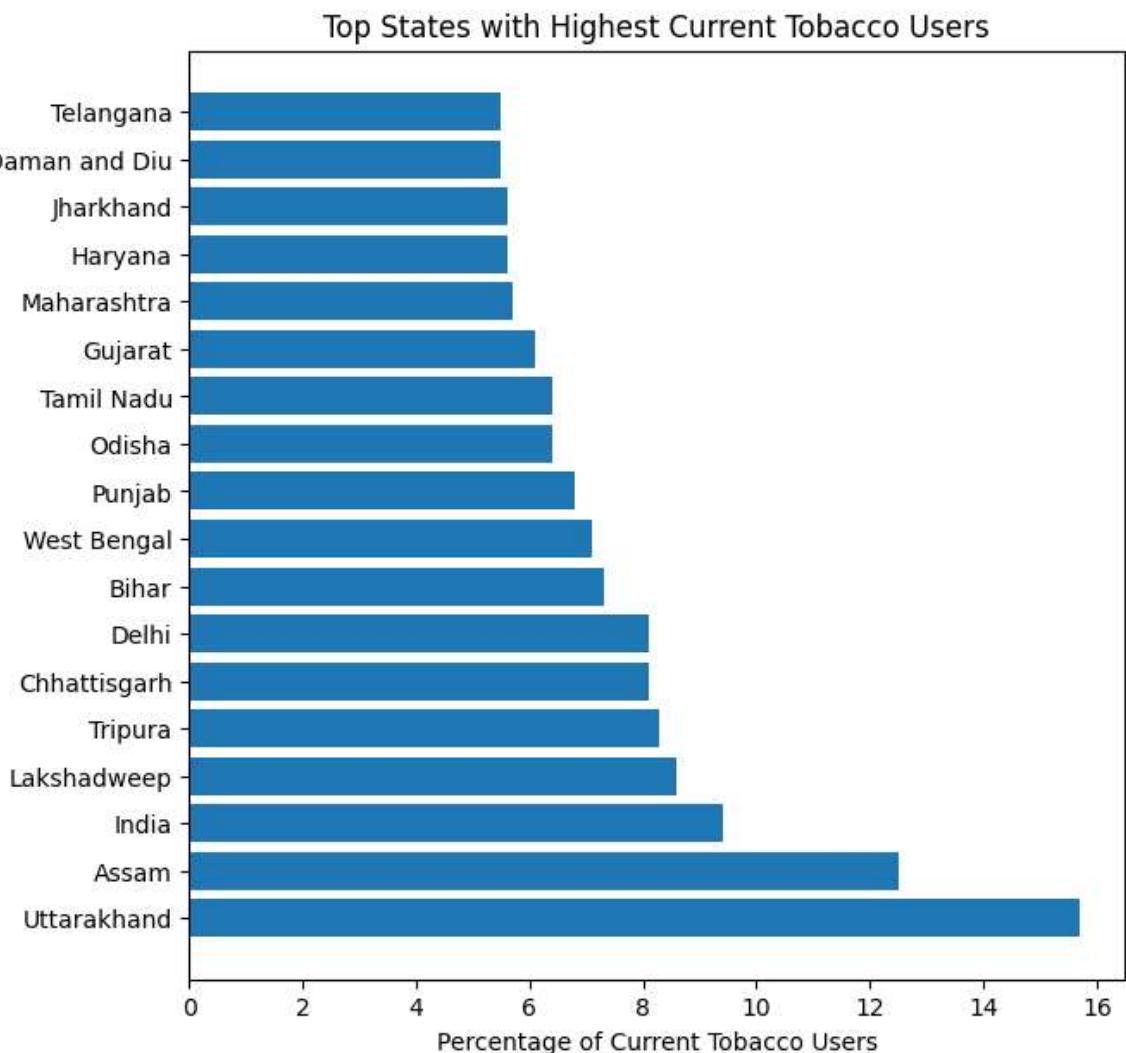
Analyzing tobacco consumption across Indian states where over 50% of the population refrains from tobacco use, underscoring prospects for effective anti-

tobacco initiatives.

Query 2

Bar Chart of Current Tobacco Users: Create a bar chart showing the "Current tobacco users (%)" for the top states.

```
In [ ]: top_current_users = data_2.nlargest(34, 'Current tobacco users (%)')
plt.figure(figsize=(7, 7))
plt.barh(top_current_users['State/UT'], top_current_users['Current tobacco users (%)'])
plt.xlabel('Percentage of Current Tobacco Users')
plt.title('Top States with Highest Current Tobacco Users')
plt.show()
```

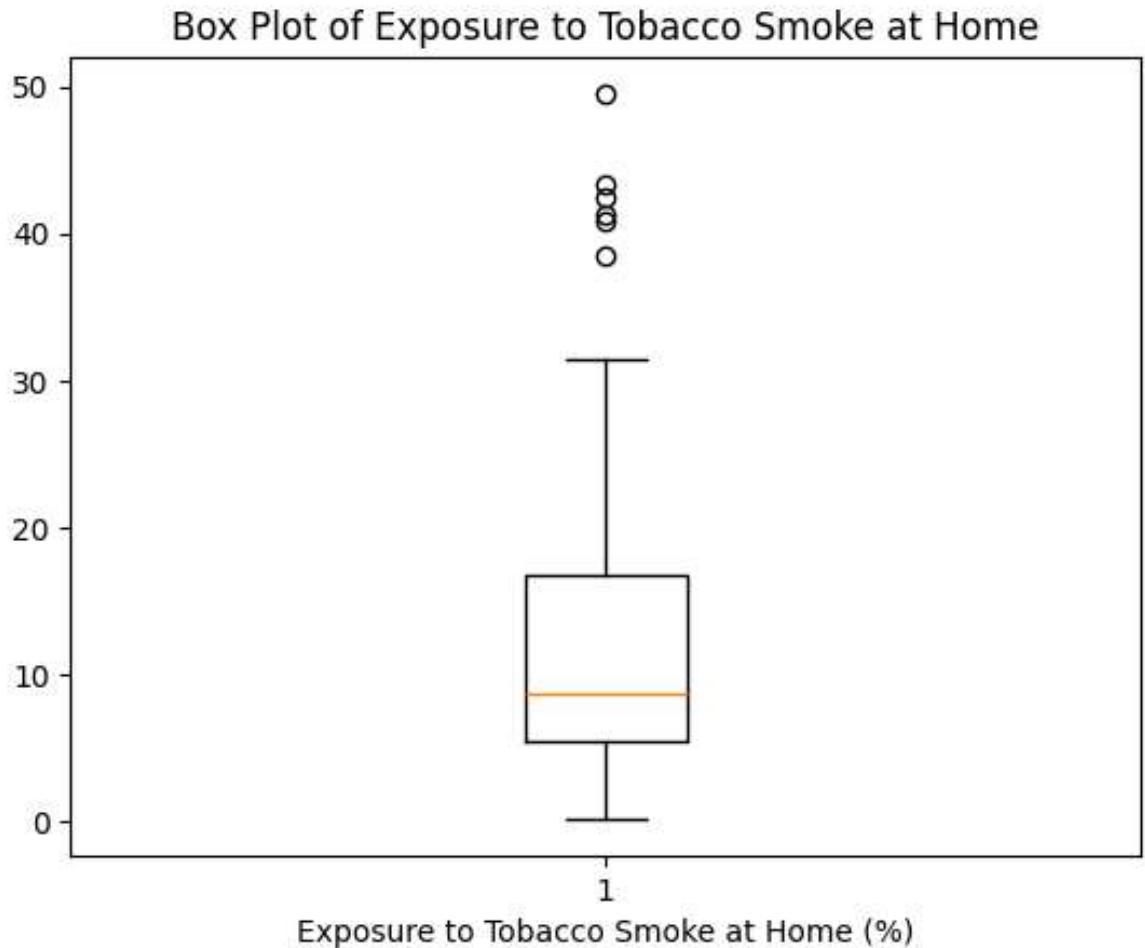


The bar chart shows different percentages of people using tobacco in the top Indian states. This tells us that there's a need for specific actions and awareness efforts to reduce tobacco use across the country.

Query 3

Box Plot of Exposure to Tobacco Smoke at Home: Create a box plot to visualize the distribution of "Exposure to tobacco smoke at home (%)".

```
In [ ]: plt.boxplot(data_1['Exposure to tobacco smoke at home (%)'])
plt.xlabel('Exposure to Tobacco Smoke at Home (%)')
plt.title('Box Plot of Exposure to Tobacco Smoke at Home')
plt.show()
```



This above query shows a box plot to depict the distribution of "Exposure to tobacco smoke at home (%)" offers a clear and concise visual representation, enabling a comprehensive understanding of the variability in this crucial health-related variable.

Query 4

Find the average percentage of people who have tried e-cigarettes across all states.

```
In [ ]: average_e_cig = data_2["Ever e-cigarette use (%)].mean()
print(f"Average % of people who tried e-cigarettes: {average_e_cig:.2f}%")
```

Average % of people who tried e-cigarettes: 2.10%

It shows average percentage of people who tried e-cigarettes. e-cigarettes is not well known in India

Query 5

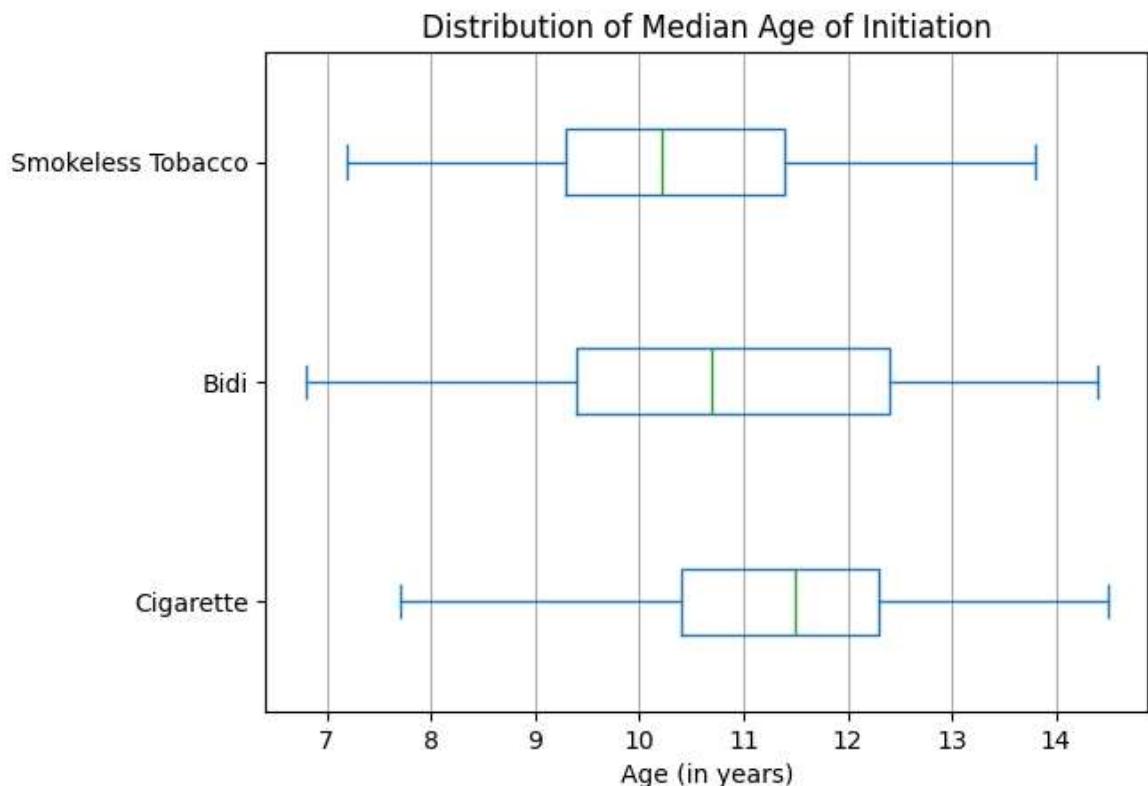
Plot the distribution of the median age of initiation for cigarettes, bidis, and smokeless tobacco.

```
In [ ]: df = data_2
columns_to_plot = [
    'Median age of initiation of Cigarette (in years)',
    'Median age of initiation of Bidi (in years)',
    'Median age of initiation of smokeless tobacco (in years)'
]

# Create a subset DataFrame with only the selected columns
initiation_df = df[columns_to_plot]

# Plot the distribution of the median age of initiation for cigarettes, bidis, and smokeless tobacco
plt.figure(figsize=(7, 6))
initiation_df.plot(kind='box', vert=False)
plt.title("Distribution of Median Age of Initiation")
plt.xlabel("Age (in years)")
plt.yticks([1, 2, 3], ['Cigarette', 'Bidi', 'Smokeless Tobacco'])
plt.grid(axis='x')
plt.show()
```

<Figure size 700x600 with 0 Axes>



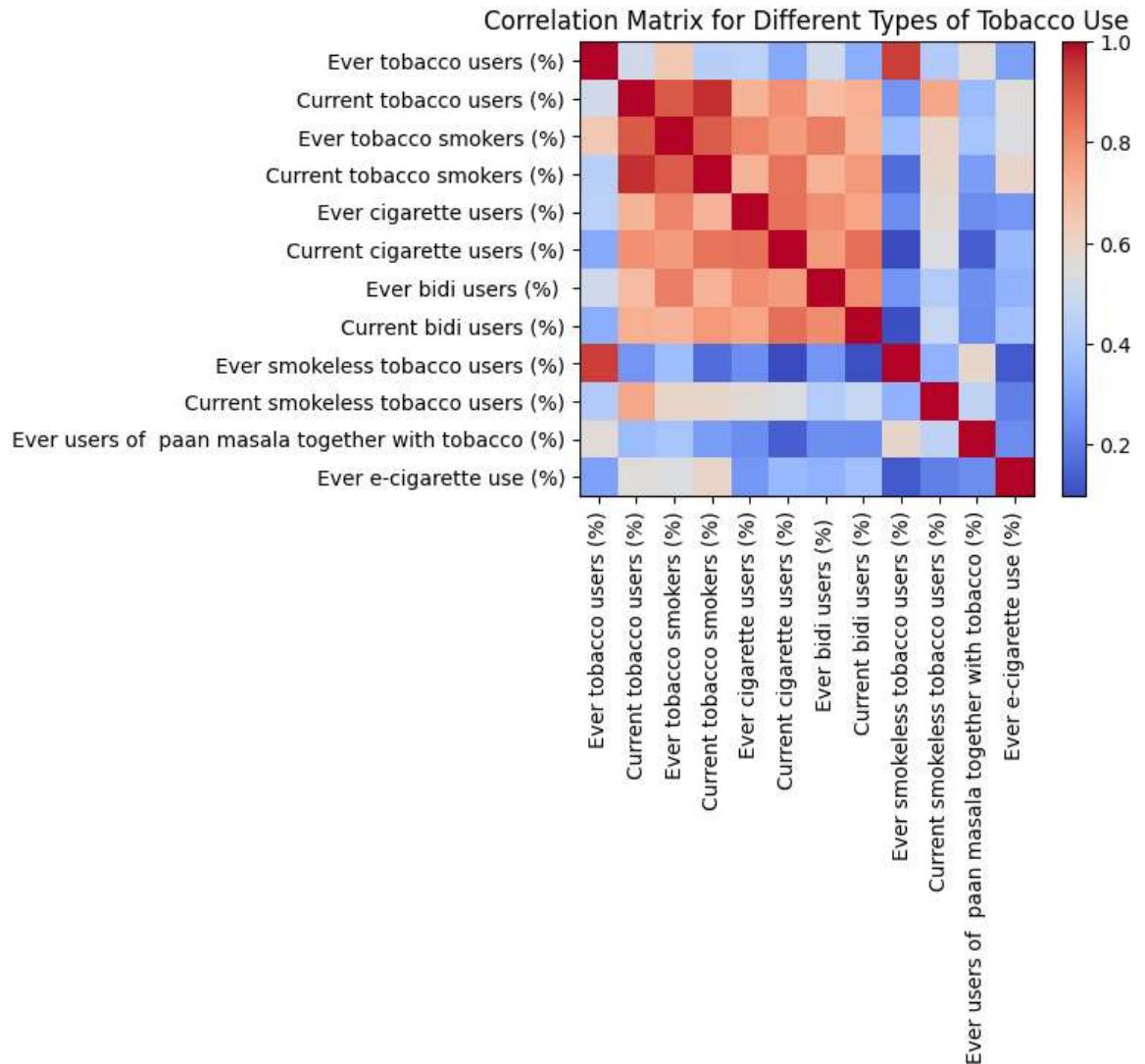
The boxplot shows distribution of the median age of initiation for cigarettes, bidis, and smokeless tobacco. Initiation of tobacco usage is too young, we can see proper action is required.

Query 6

Plot a correlation matrix for different types of tobacco use.

```
In [ ]: columns_to_correlate = [
    'Ever tobacco users (%)',
    'Current tobacco users (%)',
    'Ever tobacco smokers (%)',
    'Current tobacco smokers (%)',
    'Ever cigarette users (%)',
    'Current cigarette users (%)',
    'Ever bidi users (%)',
    'Current bidi users (%)',
    'Ever smokeless tobacco users (%)',
    'Current smokeless tobacco users (%)',
    'Ever users of paan masala together with tobacco (%)',
    'Ever e-cigarette use (%)'
]

# Create a correlation matrix and plot it as a heatmap
plt.figure(figsize=(5, 4))
corr_matrix = df[columns_to_correlate].corr()
plt.imshow(corr_matrix, cmap='coolwarm', interpolation='nearest')
plt.colorbar()
plt.title("Correlation Matrix for Different Types of Tobacco Use")
plt.xticks(range(len(columns_to_correlate)), columns_to_correlate, rotation=45)
plt.yticks(range(len(columns_to_correlate)), columns_to_correlate)
plt.show()
```



The correlation matrix for different types of tobacco use is shown. We can see tobacco, cigarette, bidi users data is highly correlated while smokeless, paan masala together with tobacco, and e-cigarette use is slightly correlated.

Query 7

Plot the percentage of current bidi and cigarette users for each state in a bar graph.

```
In [ ]: # 9. Plot the percentage of current bidi and cigarette users for each state

subset_df = df[df['Area'] == 'Rural'][['State/UT', 'Current bidi users (%)']]
subset_df.set_index('State/UT', inplace=True)
subset_df.plot(kind='bar', figsize=(12, 8), width=0.8)

plt.title("Percentage of Current Bidi and Cigarette Users by State in Rural Areas")
plt.xlabel("State/UT")
plt.ylabel("Percentage (%)")
plt.legend(title="Tobacco Use Type")

# Show the plot
plt.show()

subset_df = df[df['Area'] == 'Urban'][['State/UT', 'Current bidi users (%)']]
subset_df.set_index('State/UT', inplace=True)
subset_df.plot(kind='bar', figsize=(12, 8), width=0.8)

plt.title("Percentage of Current Bidi and Cigarette Users by State in Urban Areas")
plt.xlabel("State/UT")
plt.ylabel("Percentage (%)")
plt.legend(title="Tobacco Use Type")

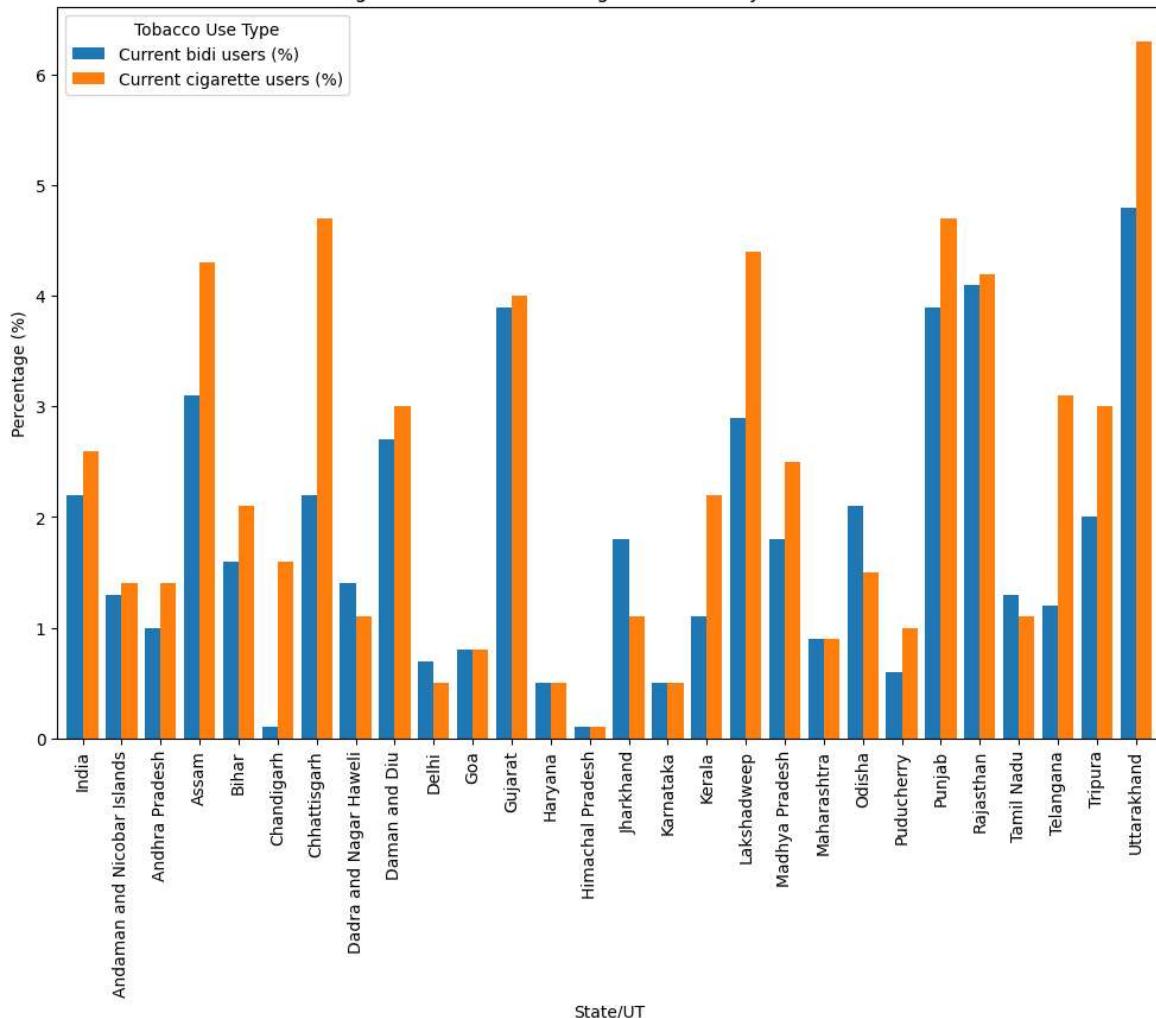
# Show the plot
plt.show()

subset_df = df[df['Area'] == 'Total'][['State/UT', 'Current bidi users (%)']]
subset_df.set_index('State/UT', inplace=True)
subset_df.plot(kind='bar', figsize=(12, 8), width=0.8)

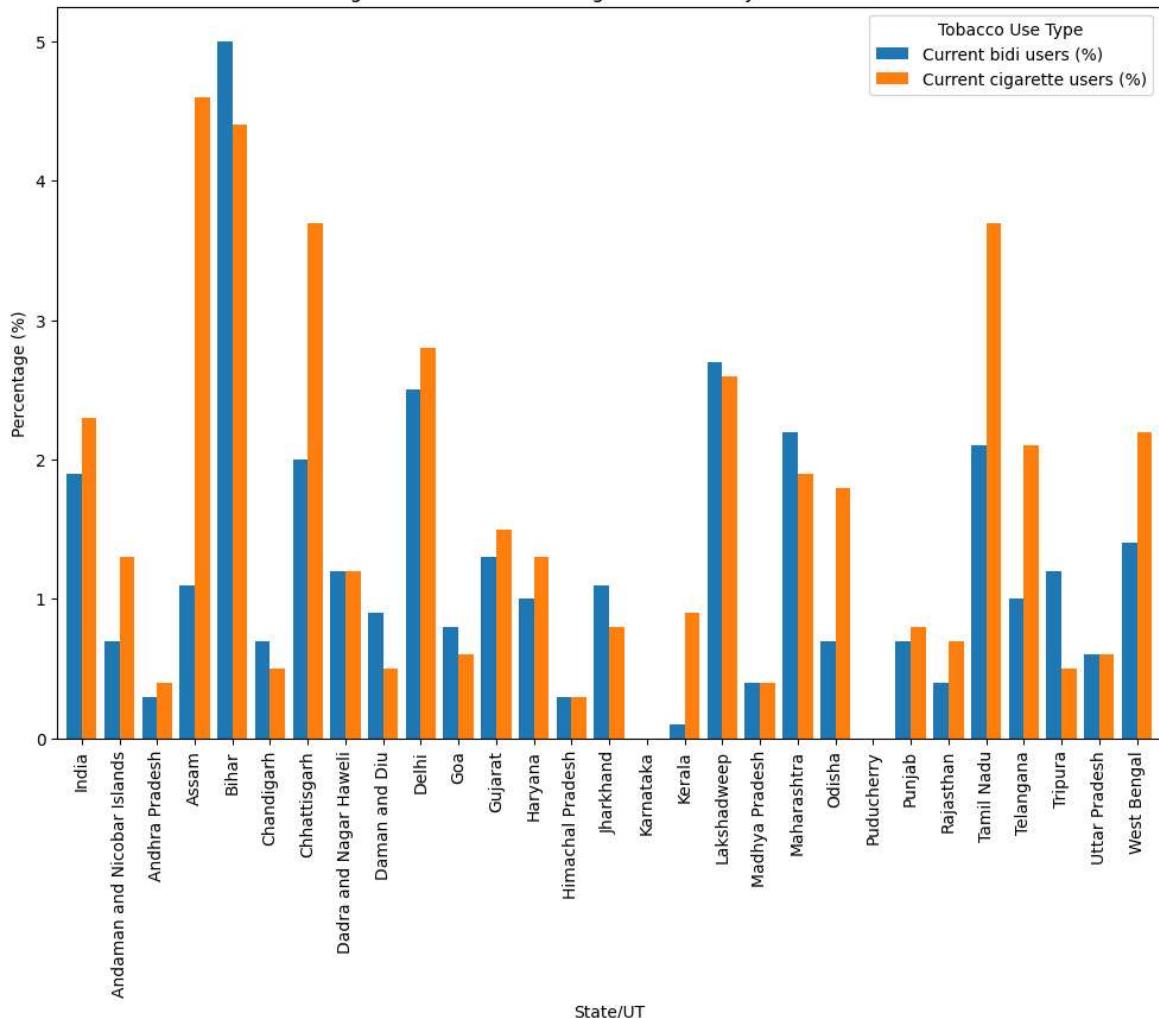
plt.title("Percentage of Current Bidi and Cigarette Users by State in Total")
plt.xlabel("State/UT")
plt.ylabel("Percentage (%)")
plt.legend(title="Tobacco Use Type")

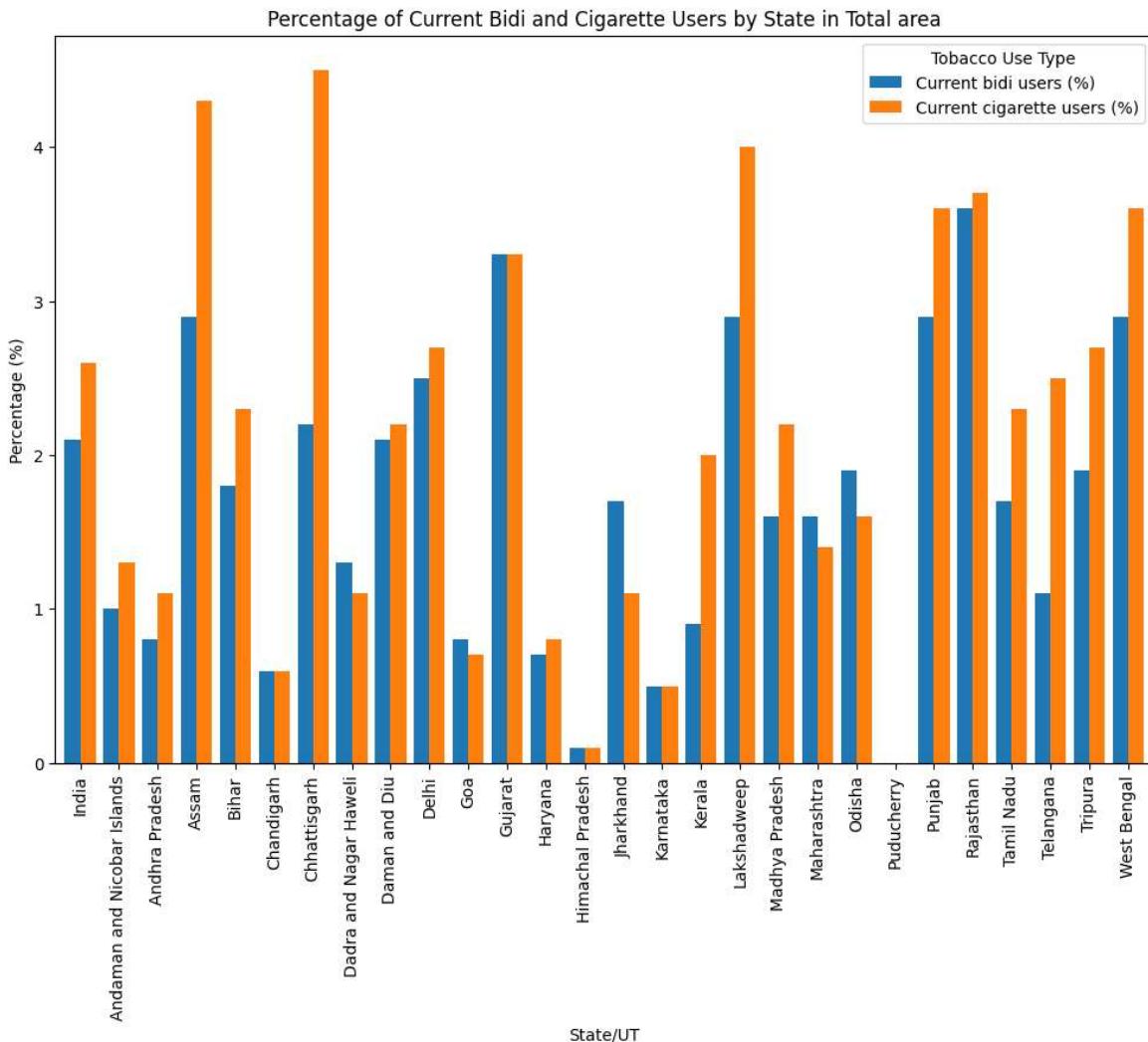
# Show the plot
plt.show()
```

Percentage of Current Bidi and Cigarette Users by State in Rural area



Percentage of Current Bidi and Cigarette Users by State in Urban area





The bar plot shows the percentage of current bidi and cigarette users for each state. We can see which states using more tobacco and less. Providing valuable information.

Query 8

Determine the state with the highest awareness about e-cigarettes.

```
In [ ]: highest_awareness_state = data_2[data_2['Awareness about e-cigarette (%)' > 0].index[0]]
print("State with the Highest Awareness about E-cigarettes:\n")
print(highest_awareness_state[['State/UT', 'Awareness about e-cigarette (%)']])
```

State with the Highest Awareness about E-cigarettes:

State/UT	Awareness about e-cigarette (%)	
97	Tripura	44.9

The above query shows the state which has highest awareness about e-cigarettes.

Query 9

Show states where the gap between ever tobacco users and current tobacco users is more than 20%.

```
In [ ]: states_with_gap_gt_20 = data_2[abs(data_2['Ever tobacco users (%)'] - data_2['Current tobacco users (%)']) > 20]
print("States with a Gap of More than 20% between Ever and Current Tobacco Users:")
print(states_with_gap_gt_20[['State/UT', 'Ever tobacco users (%)', 'Current tobacco users (%)']])
```

States with a Gap of More than 20% between Ever and Current Tobacco Users:

	State/UT	Ever tobacco users (%)	\
5	Andaman and Nicobar Islands		26.8
42	Himachal Pradesh		33.4
44	Himachal Pradesh		36.3
47	Jharkhand		27.0

	Current tobacco users (%)
5	5.1
42	1.0
44	1.0
47	5.6

The above query reveals states that has 20% gap between ever tobacco users and current users. These states are stopping smoking, we can do research how these states are quitting smoking.

Query 10

Display the top 5 states with the highest percentage of current tobacco users.

```
In [ ]: top_states = data_2[data_2["Area"]=="Total"].sort_values(by="Current tobacco users (%)", ascending=False)
print(top_states[['State/UT', "Current tobacco users (%)"]])
```

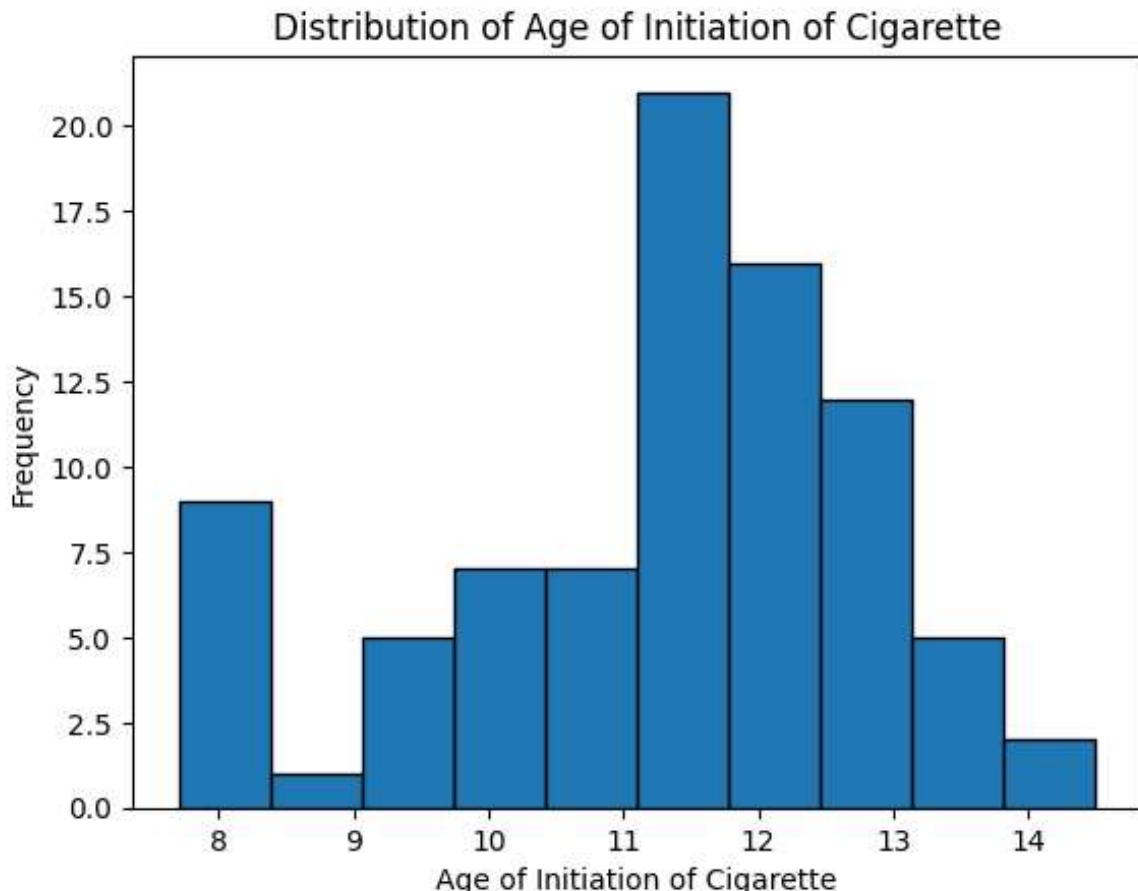
	State/UT	Current tobacco users (%)
12	Assam	11.9
0	India	8.5
54	Lakshadweep	8.1
30	Delhi	8.0
21	Chhattisgarh	8.0

The above query reveals the top 5 states exhibiting the highest percentage of current tobacco users, providing valuable insights for targeted public health interventions and awareness campaigns in these regions.

Query 11

Histogram of Age of Initiation of Cigarette: Plot a histogram showing the distribution of the "Median age of initiation of Cigarette (in years)".

```
In [ ]: plt.hist(data_2['Median age of initiation of Cigarette (in years)'].dropna()
plt.xlabel('Age of Initiation of Cigarette')
plt.ylabel('Frequency')
plt.title('Distribution of Age of Initiation of Cigarette')
plt.show()
```



The above query visualize the histogram of the median age of initiation of cigarettes provides a comprehensive and insightful representation of the distribution, offering valuable insights into the patterns and trends associated with the age at which individuals typically start smoking.

Query 12

Find states where more than 60% of the current tobacco users want to quit now.

```
In [ ]: quit_intent = data_2[data_2["Area"]=="Total"]
quit_intent_1 = quit_intent[quit_intent["Current tobacco smokers who wanted to q
print(quit_intent_1[["State/UT", "Current tobacco smokers who wanted to q
```

State/UT		Current tobacco smokers who wanted to quit smoking now (%)
48	Karnataka	99.9
84	Rajasthan	88.9

The above query is identifying states with over 60% of current tobacco users expressing a desire to quit emphasizes the urgent need for targeted intervention

and support programs to address and reduce tobacco dependence in those regions.

Query 13

Show the top 5 states where people are most susceptible to future cigarette use.

```
In [ ]: susceptible_states = data_2[data_2["Area"]=="Total"].sort_values(by="Never cigarette smokers susceptible to cigarette use in future (%)", ascending=False).head(5)
```

	State/UT	Never cigarette smokers susceptible to cigarette use in future (%)
84	Rajasthan	17.1
75	Odisha	13.9
15	Bihar	13.7
30	Delhi	13.2
3	Andaman and Nicobar Islands	10.2

This query reveals the top 5 states with the highest susceptibility to future cigarette use indicate a concerning trend that warrants targeted intervention and prevention efforts to address potential public health risks.

Query 14

Show states where exposure to tobacco smoke at home is more than in public places.

```
In [ ]: high_home_exposure = data_2[data_2["Exposure to tobacco smoke at home (%)"] > data_2["Exposure to tobacco smoke in public places (%)", ascending=False].head(1)
```

	State/UT
55	Lakshadweep

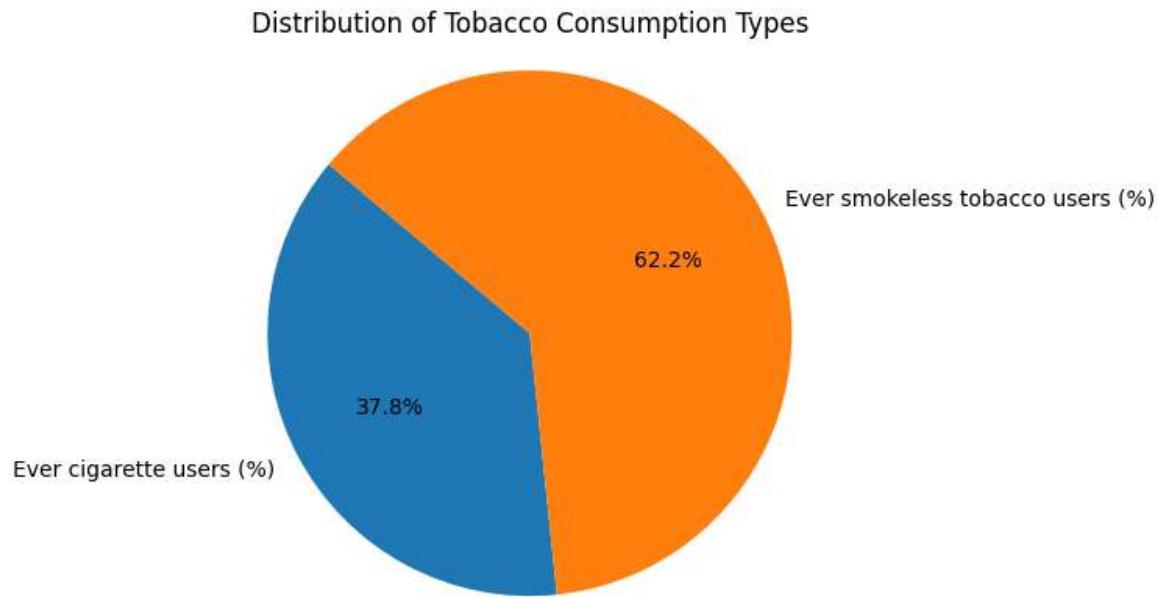
The query reveals Indian states where home tobacco exposure surpasses public places. This highlights potential disparities in smoking regulations and behavioral patterns. Targeted initiatives could mitigate health risks, focusing on reducing indoor exposure and promoting awareness in regions with this disparity for improved public health outcomes.

Query 15

Pie Chart of Tobacco Consumption Types: Show a pie chart representing the breakdown of "Ever tobacco users (%)" into cigarette and smokeless tobacco

users.

```
In [ ]: tobacco_types = ['Ever cigarette users (%)', 'Ever smokeless tobacco users (%)']
consumption_percentages = data_1[tobacco_types].sum()
plt.pie(consumption_percentages, labels=tobacco_types, autopct='%1.1f%%',
plt.axis('equal')
plt.title('Distribution of Tobacco Consumption Types')
plt.show()
```



The pie chart illustrates the proportion of ever tobacco users between cigarettes and smokeless tobacco. It identifies the dominant type, aiding in understanding their relative prevalence and suggesting potential targeted interventions based on usage patterns.