Name: Shivam Bhatt

Enrollment No: 92301733046

!pip install kaggle

import os

os.environ['KAGGLE\_CONFIG\_DIR']="/content"

!kaggle datasets download stackoverflow/stack-overflow-2018-developer-survey

!unzip /content/stack-overflow-2018-developer-survey.zip

#import basic liabraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

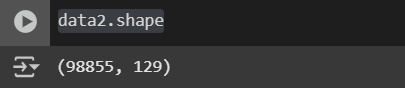
import seaborn as sns

data2=pd.read\_csv("/content/survey\_results\_public.csv")

data3=pd.read\_csv("/content/survey\_results\_schema.csv")

data2.shape

output:

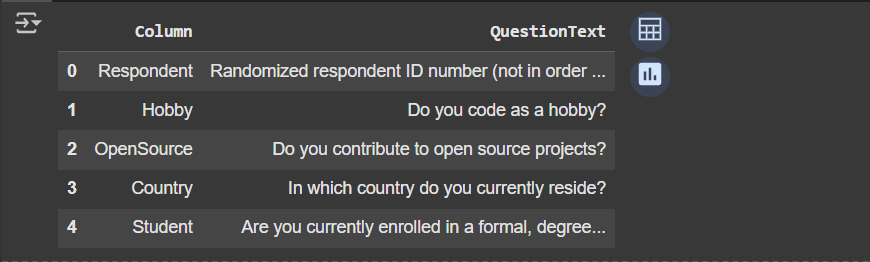


data3.shape

output:



data3.head()

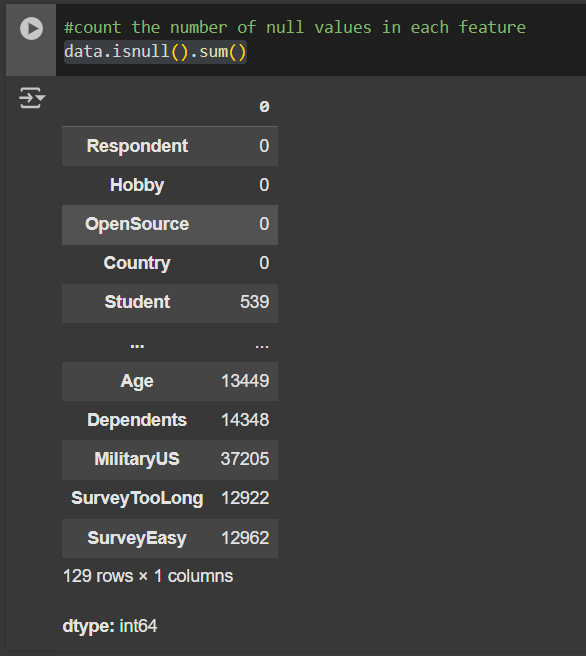


**#count the number of null values in each feature**

data=pd.read\_csv("/content/survey\_results\_public.csv")

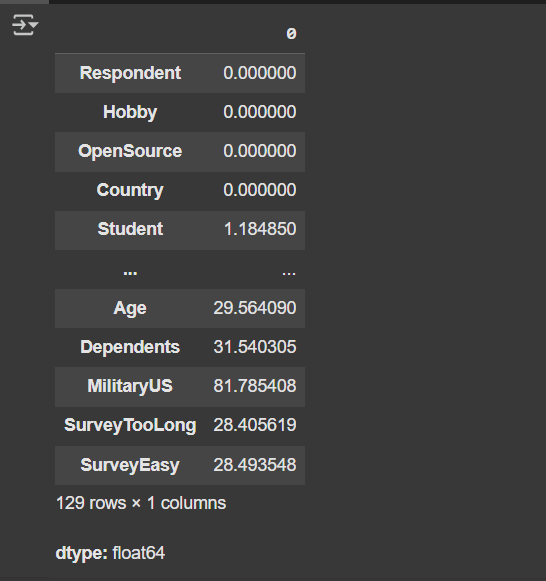
print(data)

data.isnull().sum()



**#count the percentage of null values**

data.isnull().sum()/data.shape[0]\*100



**#draw the pychart for number of people who finds coding as hobby**

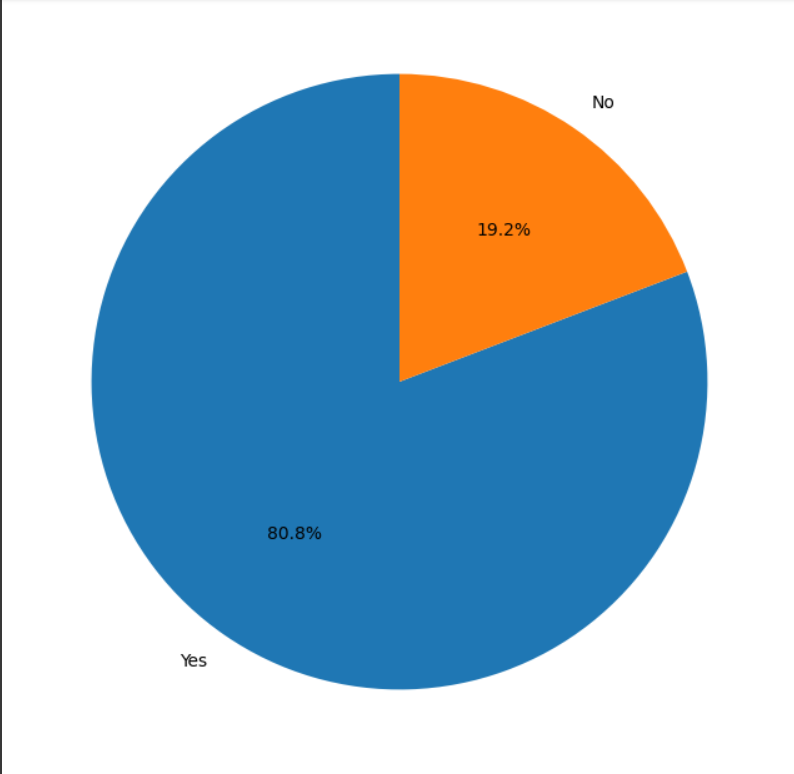
hobby\_counts = data2['Hobby'].value\_counts()

plt.figure(figsize=(8, 8))

plt.pie(hobby\_counts, labels=hobby\_counts.index, autopct='%1.1f%%', startangle=90)

plt.title('Distribution of people who code as a hobby')

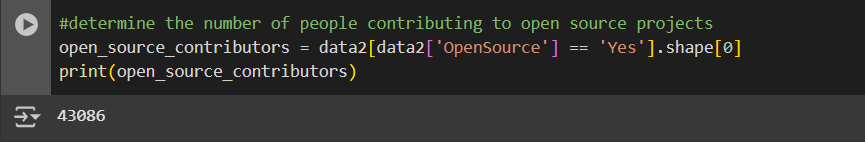
plt.show()



**#determine the number of people contributing to open source projects**

open\_source\_contributors = data2[data2['OpenSource'] == 'Yes'].shape[0]

print(open\_source\_contributors)

****

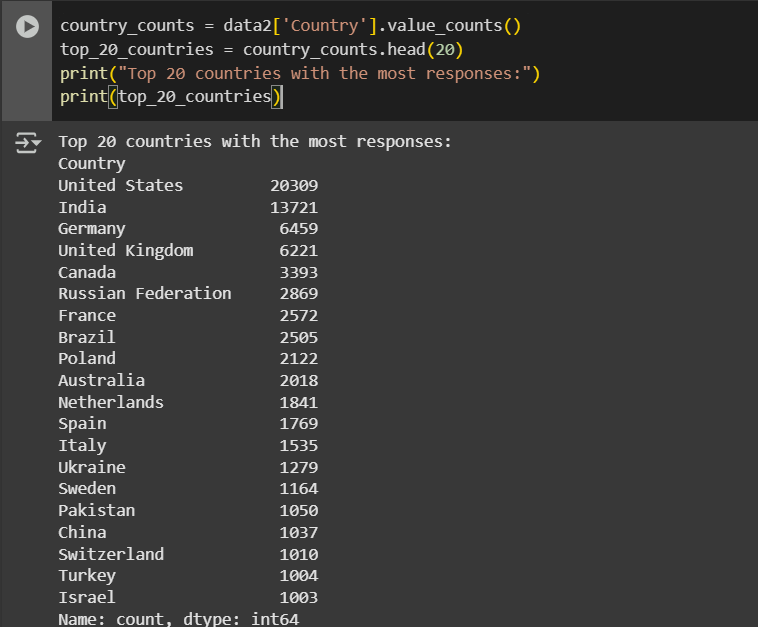
**#determine the top 20 countries where the responses are obtained**

country\_counts = data2['Country'].value\_counts()

top\_20\_countries = country\_counts.head(20)

print("Top 20 countries with the most responses:")

print(top\_20\_countries)



**#do other 5 analysis as per your own thinking (which involves different charts and graphs)**

# 1. Distribution of 'Student' status

plt.figure(figsize=(8, 6))

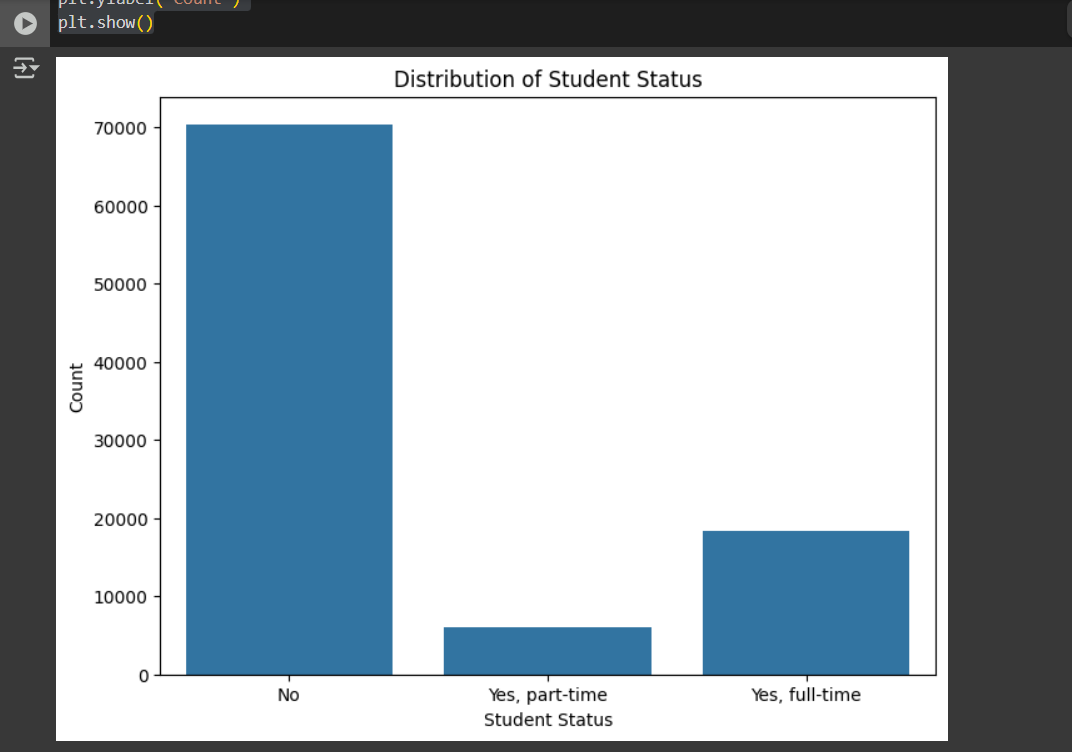
sns.countplot(data=data2, x='Student')

plt.title('Distribution of Student Status')

plt.xlabel('Student Status')

plt.ylabel('Count')

plt.show()



# 2. Distribution of 'FormalEducation'

plt.figure(figsize=(10, 6))

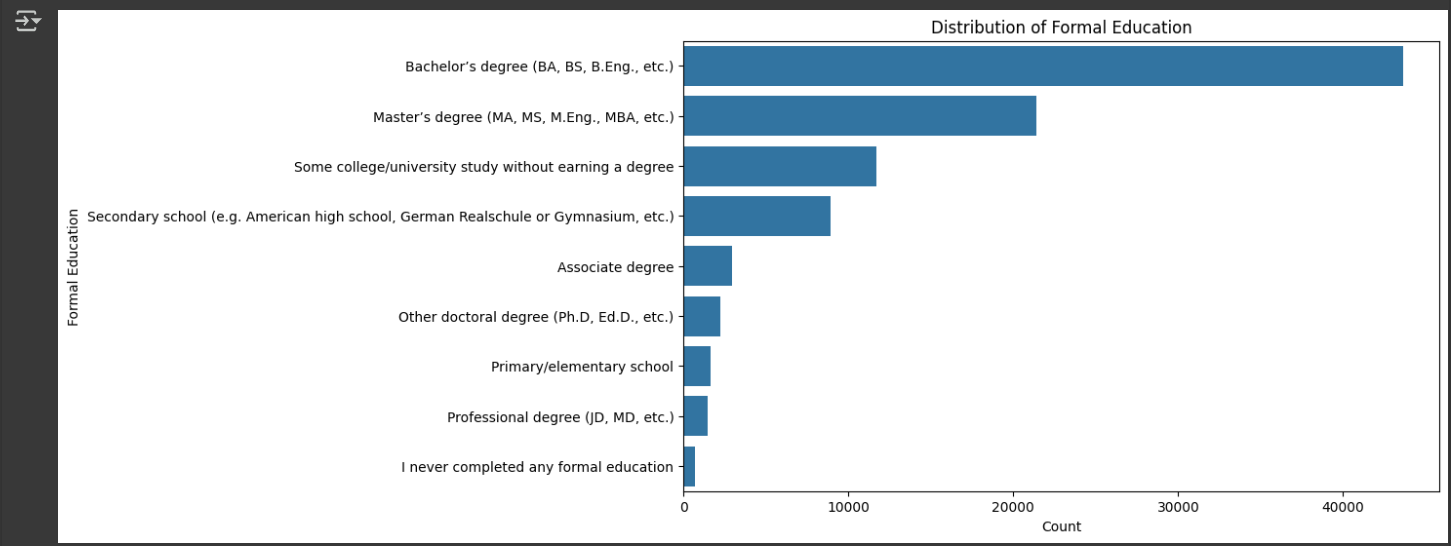
sns.countplot(data=data2, y='FormalEducation', order=data2['FormalEducation'].value\_counts().index)

plt.title('Distribution of Formal Education')

plt.xlabel('Count')

plt.ylabel('Formal Education')

plt.show()



# 3. Distribution of 'Age'

plt.figure(figsize=(10, 6))

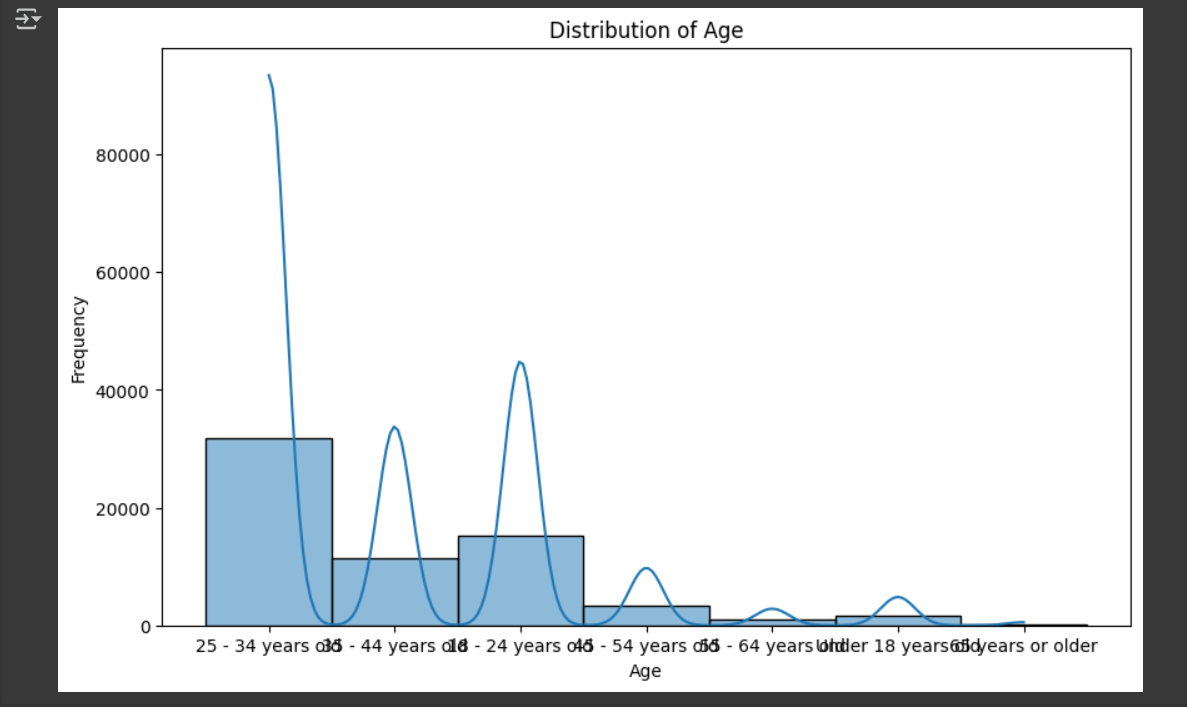
sns.histplot(data=data2, x='Age', kde=True)

plt.title('Distribution of Age')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()



# 4. Distribution of 'DevType' (Developer Type)

dev\_types = data2['DevType'].str.split(';').explode().str.strip()

dev\_type\_counts = dev\_types.value\_counts().head(15) # Displaying top 15 for clarity

plt.figure(figsize=(12, 8))

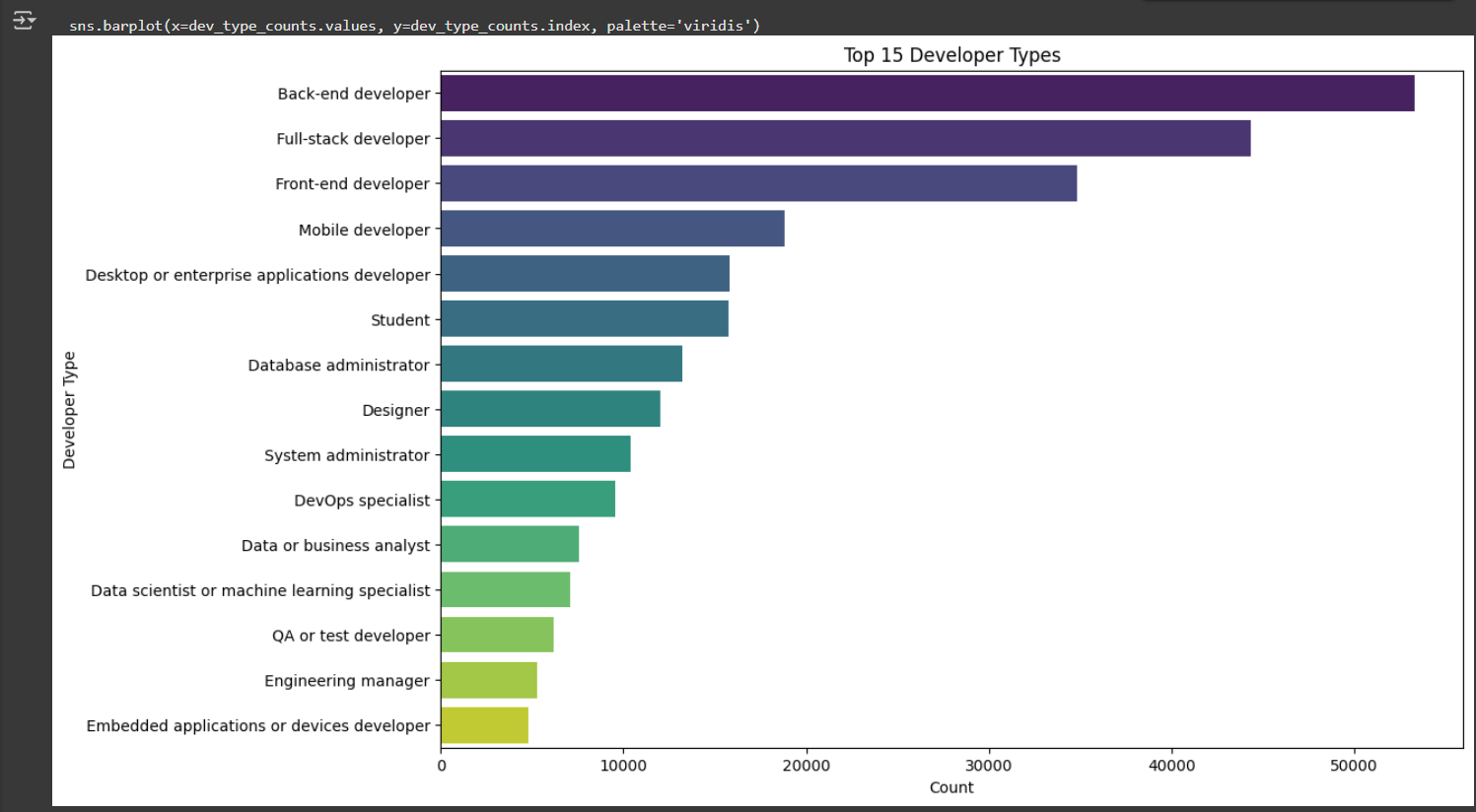
sns.barplot(x=dev\_type\_counts.values, y=dev\_type\_counts.index, palette='viridis')

plt.title('Top 15 Developer Types')

plt.xlabel('Count')

plt.ylabel('Developer Type')

plt.show()



# 5. Relationship between 'Country' and 'Hobby' for top countries

# Get top N countries (e.g., top 10)

top\_countries\_list = top\_20\_countries.head(10).index.tolist()

data\_top\_countries = data2[data2['Country'].isin(top\_countries\_list)]

hobby\_country\_counts = pd.crosstab(data\_top\_countries['Country'], data\_top\_countries['Hobby'])

hobby\_country\_counts.plot(kind='bar', stacked=True, figsize=(12, 7))

plt.title('Hobby Status by Top 10 Countries')

plt.xlabel('Country')

plt.ylabel('Number of Respondents')

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()

