The analysis of UK smoking data aims to examine patterns and trends in smoking prevalence. By analyzing factors such as age, gender, and Education Qualification, the study aims to gain insights into smoking behaviors.

In [1]:

```
# Importing Neccessary Libraries
```

In [2]:

```
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
```

In [3]:

```
# reading csv file
```

In [4]:

```
df = pd.read_csv('C:/Users/sanjith/Desktop/smoking.csv', encoding="unicode_escape")
```

In [5]:

```
df.head(10)
```

Out[5]:

	Unnamed: 0	gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_income
0	1	Male	38	Divorced	No Qualification	British	White	2,600 to 5,200
1	2	Female	42	Single	No Qualification	British	White	Under 2,600
2	3	Male	40	Married	Degree	English	White	28,600 to 36,400
3	4	Female	40	Married	Degree	English	White	10,400 to 15,600
4	5	Female	39	Married	GCSE/O Level	British	White	2,600 to 5,200
5	6	Female	37	Married	GCSE/O Level	British	White	15,600 to 20,800
6	7	Male	53	Married	Degree	British	White	Above 36,400
7	8	Male	44	Single	Degree	English	White	10,400 to 15,600
8	9	Male	40	Single	GCSE/CSE	English	White	2,600 to 5,200
9	10	Female	41	Married	No Qualification	English	White	5,200 to 10,400

In [6]:

```
df.tail(10)
```

Out[6]:

	Unnamed: 0	gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_incon
1681	1682	Male	53	Single	No Qualification	Scottish	White	20,800 28,6
1682	1683	Female	63	Married	No Qualification	British	White	Refus
1683	1684	Male	35	Married	No Qualification	Scottish	White	10,400 15,6
1684	1685	Male	78	Widowed	No Qualification	Scottish	White	Refus
1685	1686	Female	31	Single	Other/Sub Degree	Scottish	White	5,200 10,4
1686	1687	Male	22	Single	No Qualification	Scottish	White	2,600 to 5,2
1687	1688	Female	49	Divorced	Other/Sub Degree	English	White	2,600 to 5,2
1688	1689	Male	45	Married	Other/Sub Degree	Scottish	White	5,200 10,4
1689	1690	Female	51	Married	No Qualification	English	White	2,600 to 5,2
1690	1691	Male	31	Married	Degree	Scottish	White	10,400 15,6

In [7]:
```
df.shape
```
Out[7]:
```
(1691, 13)
```

In [8]:
```
df.columns
```
Out[8]:
```
Index(['Unnamed: 0', 'gender', 'age', 'marital_status',
       'highest_qualification', 'nationality', 'ethnicity', 'gross_income',
       'region', 'smoke', 'amt_weekends', 'amt_weekdays', 'type'],
      dtype='object')
```

In [9]:
```
# Data Inquiry Questions :

# 1) What is the distribution of smoking prevalence among different genders in the dat
# 2) How does smoking behavior vary across different age groups?
# 3) Is there any correlation between marital status and smoking habits ?
# 4) What is the relation between the highest level of education attained and smoking
# 5) Are there any regional differences in smoking prevalence in the UK?
```

In [10]:
```
# Dropping Unneccessary columns
```

In [11]:
```
drop_columns = ['ethnicity', 'gross_income', 'region', 'amt_weekdays', 'type','Unnamec

df.drop(drop_columns, inplace=True, axis= 1)
```

In [12]:
```
df.head(5)
```

Out[12]:

	gender	age	marital_status	highest_qualification	nationality	smoke
0	Male	38	Divorced	No Qualification	British	No
1	Female	42	Single	No Qualification	British	Yes
2	Male	40	Married	Degree	English	No
3	Female	40	Married	Degree	English	No
4	Female	39	Married	GCSE/O Level	British	No

In [13]:

```
    df= df.rename(columns ={ 'gender':'Gender' , 'age':'Age', 'marital_status':'Marita
```

In [14]:

```
df.head(10)
```

Out[14]:

	Gender	Age	Marital Status	Education	Nationality	Status
0	Male	38	Divorced	No Qualification	British	No
1	Female	42	Single	No Qualification	British	Yes
2	Male	40	Married	Degree	English	No
3	Female	40	Married	Degree	English	No
4	Female	39	Married	GCSE/O Level	British	No
5	Female	37	Married	GCSE/O Level	British	No
6	Male	53	Married	Degree	British	Yes
7	Male	44	Single	Degree	English	No
8	Male	40	Single	GCSE/CSE	English	Yes
9	Female	41	Married	No Qualification	English	Yes

In [15]:

```
df.isna().sum()
```

Out[15]:

```
Gender            0
Age               0
Marital Status    0
Education         0
Nationality       0
Status            0
dtype: int64
```

In [16]:

```
df.duplicated()
```

Out[16]:

```
0       False
1       False
2       False
3       False
4       False
        ...  
1686    False
1687    False
1688    False
1689     True
1690    False
Length: 1691, dtype: bool
```

In [17]:

```
df.drop_duplicates()
```

Out[17]:

	Gender	Age	Marital Status	Education	Nationality	Status
0	Male	38	Divorced	No Qualification	British	No
1	Female	42	Single	No Qualification	British	Yes
2	Male	40	Married	Degree	English	No
3	Female	40	Married	Degree	English	No
4	Female	39	Married	GCSE/O Level	British	No
...	...	...	...	...	...	...
1685	Female	31	Single	Other/Sub Degree	Scottish	No
1686	Male	22	Single	No Qualification	Scottish	No
1687	Female	49	Divorced	Other/Sub Degree	English	Yes
1688	Male	45	Married	Other/Sub Degree	Scottish	No
1690	Male	31	Married	Degree	Scottish	No

1436 rows × 6 columns

In [18]:

```
# # 1) What is the distribution of smoking prevalence among different genders in the c
```

In [19]:

```
 df['Gender'].unique()
```

Out[19]:

```
array(['Male', 'Female'], dtype=object)
```

In [20]:

```
df['Gender'].value_counts()
```

Out[20]:

```
Female    965
Male      726
Name: Gender, dtype: int64
```

In [21]:

```
df['Status'].value_counts()
```

Out[21]:

```
No     1270
Yes     421
Name: Status, dtype: int64
```

In [22]:

```

total_male = len(df[(df['Gender'] == 'Male') & (df['Status'] == 'Yes')])
```

```
total_female = len(df[(df['Gender'] == 'Female') & (df['Status'] == 'Yes')])

print( total_male)
print ( total_female)
```

```
187
234
```

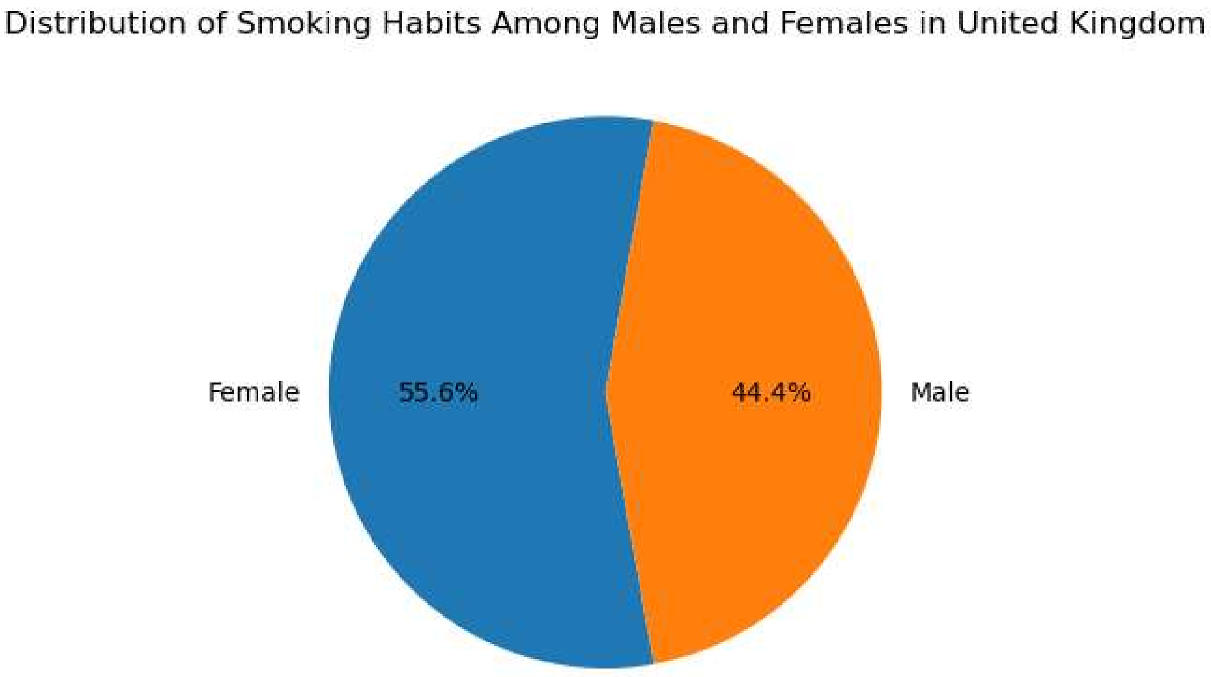
In [23]:

```
gender_counts = [234, 187]
gender_labels = ['Female', 'Male']


plt.pie(gender_counts, labels=gender_labels, autopct='%1.1f%%', startangle=80)


plt.title('Distribution of Smoking Habits Among Males and Females in United Kingdom')


plt.show()
```



In [24]:

```
# 2) How does smoking behavior vary across different age groups?
```

In [25]:

```
df['Age'] = df['Age'].astype(float).astype(int)

bins = [10, 20, 30, 40, 50, 60, 70, 80]
labels = ['10-20', '20-30', '30-40', '40-50', '50-60','60-70','70-80']

df['Age'] = pd.cut(df['Age'], bins=bins, labels=labels)
```

In [26]:

```
df.head(5)
```

Out[26]:

	Gender	Age	Marital Status	Education	Nationality	Status
0	Male	30-40	Divorced	No Qualification	British	No
1	Female	40-50	Single	No Qualification	British	Yes
2	Male	30-40	Married	Degree	English	No
3	Female	30-40	Married	Degree	English	No
4	Female	30-40	Married	GCSE/O Level	British	No

In [27]:

```
smokers_by_age_group = df[df['Status'] == 'Yes'].groupby('Age').size()
print(smokers_by_age_group)
```

```
Age
10-20     22
20-30     87
30-40    104
40-50     85
50-60     55
60-70     39
70-80     25
dtype: int64
```

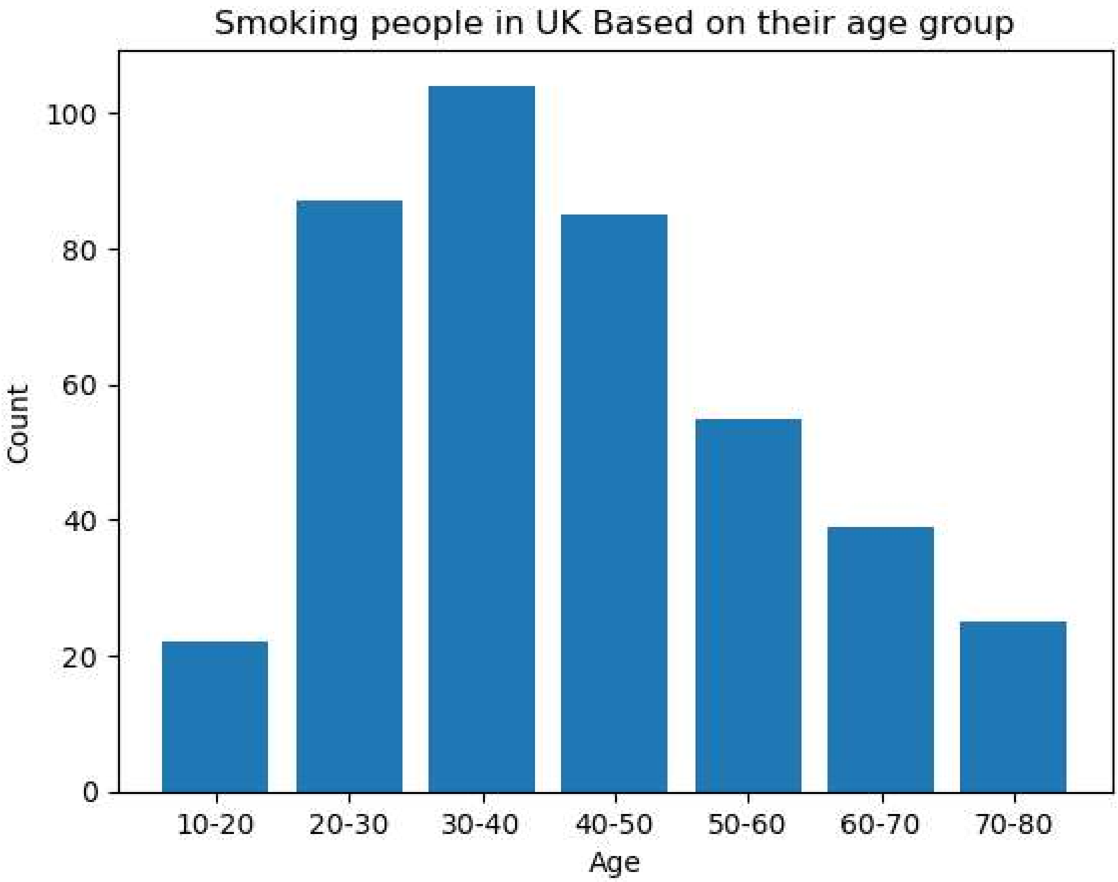
In [28]:

```

plt.bar(smokers_by_age_group.index, smokers_by_age_group.values)
plt.xlabel('Age')
plt.ylabel('Count')
plt.title('Smoking people in UK Based on their age group')



plt.show()
```



In [29]:

```
non_smokers_by_age_group = df[df['Status'] == 'No'].groupby('Age').size()
print(non_smokers_by_age_group)
```

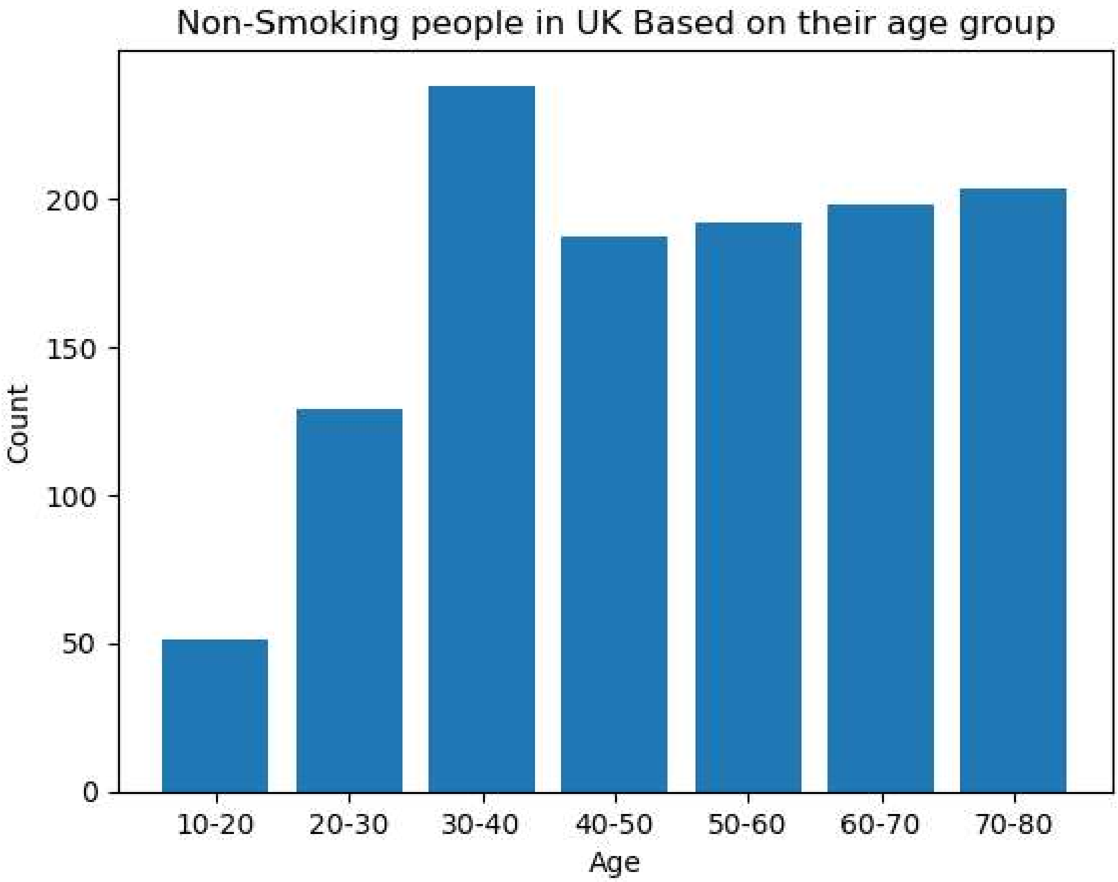
```
Age
10-20     51
20-30    129
30-40    238
40-50    187
50-60    192
60-70    198
70-80    203
dtype: int64
```

In [30]:

```

plt.bar(non_smokers_by_age_group.index, non_smokers_by_age_group.values )
plt.xlabel('Age')
plt.ylabel('Count')
plt.title('Non-Smoking people in UK Based on their age group')


plt.show()
```



In [31]:

```
# 3) Is there any correlation between marital status and smoking habits ?
```

In [32]:

```
df['Marital Status'].value_counts()
```

Out[32]:

```
Married      812
Single       427
Widowed      223
Divorced     161
Separated     68
Name: Marital Status, dtype: int64
```

In [33]:

```
smoking_married = len(df[(df['Marital Status'] == 'Married') & (df['Status'] == 'Yes')

smoking_single = len(df[(df['Marital Status'] == 'Single') & (df['Status'] == 'Yes')])
smoking_widowed = len(df[(df['Marital Status'] == 'Widowed') & (df['Status'] == 'Yes')
smoking_divorced = len(df[(df['Marital Status'] == 'Divorced') & (df['Status'] == 'Yes
smoking_separated = len(df[(df['Marital Status'] == 'Separated') & (df['Status'] == 'Y

print (smoking_married)
print(smoking_single)
print(smoking_widowed)
print(smoking_divorced)
print(smoking_separated)
```

```
143
158
40
58
22
```

In [34]:

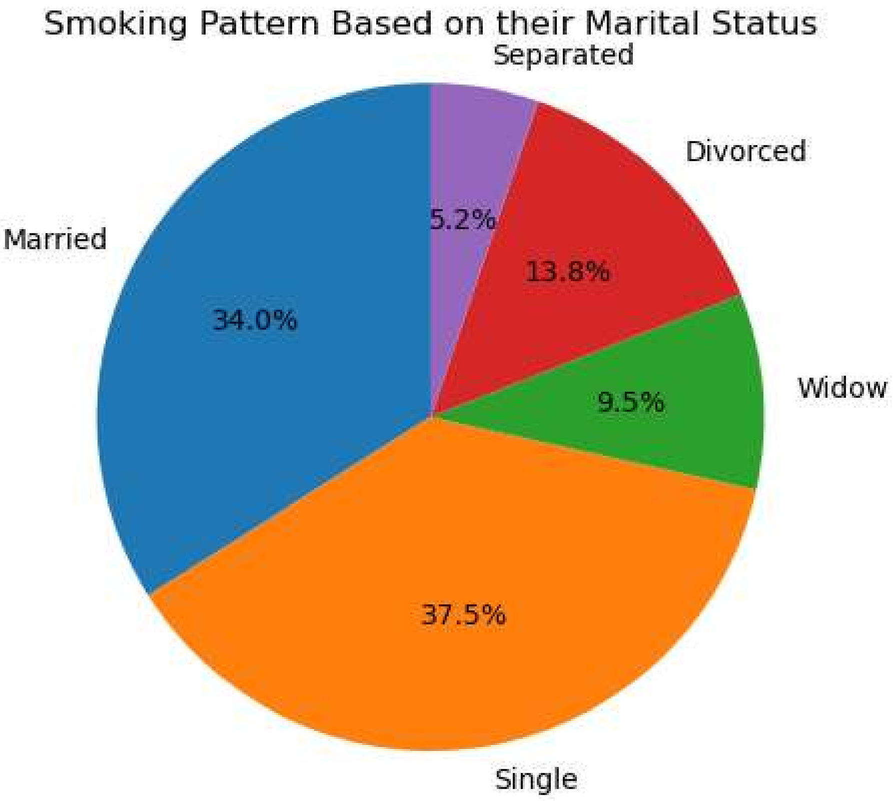
```
labels = ['Married', 'Single', 'Widow', 'Divorced', 'Separated']
sizes = [smoking_married, smoking_single, smoking_widowed, smoking_divorced, smoking_s


plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)


plt.title('Smoking Pattern Based on their Marital Status')


plt.axis('equal')


plt.show()
```



In [35]:

```
# What is the relation between the highest level of education attained and smoking rat
```

In [36]:

```
df['Education'].value_counts()
```

Out[36]:

```
No Qualification     586
GCSE/O Level         308
Degree               262
Other/Sub Degree     127
Higher/Sub Degree    125
A Levels             105
GCSE/CSE             102
ONC/BTEC              76
Name: Education, dtype: int64
```

In [37]:

```
df[df['Status'] == 'Yes']['Education'].value_counts()
```

Out[37]:

```
No Qualification     137
GCSE/O Level         105
Degree                39
GCSE/CSE              38
Other/Sub Degree      31
Higher/Sub Degree     27
ONC/BTEC              23
A Levels              21
Name: Education, dtype: int64
```

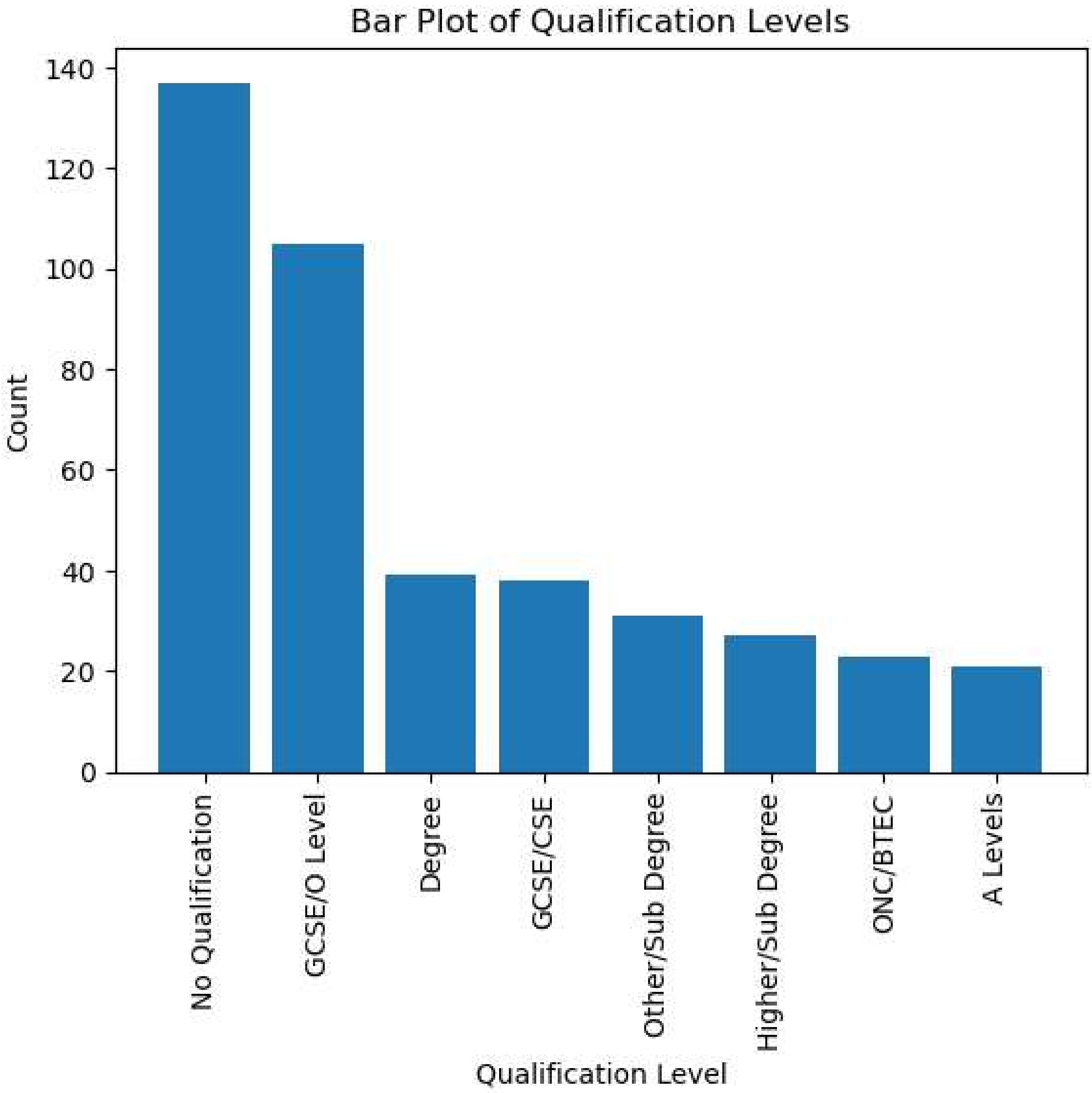
In [38]:

```
import matplotlib.pyplot as plt

qualification = ['No Qualification', 'GCSE/O Level', 'Degree', 'GCSE/CSE', 'Other/Sub
count = [137, 105, 39, 38, 31, 27, 23, 21]

plt.bar(qualification, count)
plt.xlabel('Qualification Level')
plt.ylabel('Count')
plt.title('Bar Plot of Qualification Levels')
plt.xticks(rotation=90)

plt.show()
```



```
In [39]: # 5) Are there any regional differences in smoking prevalence in the UK?
```

```
In [40]: df[df['Status'] == 'Yes']['Nationality'].value_counts()
```

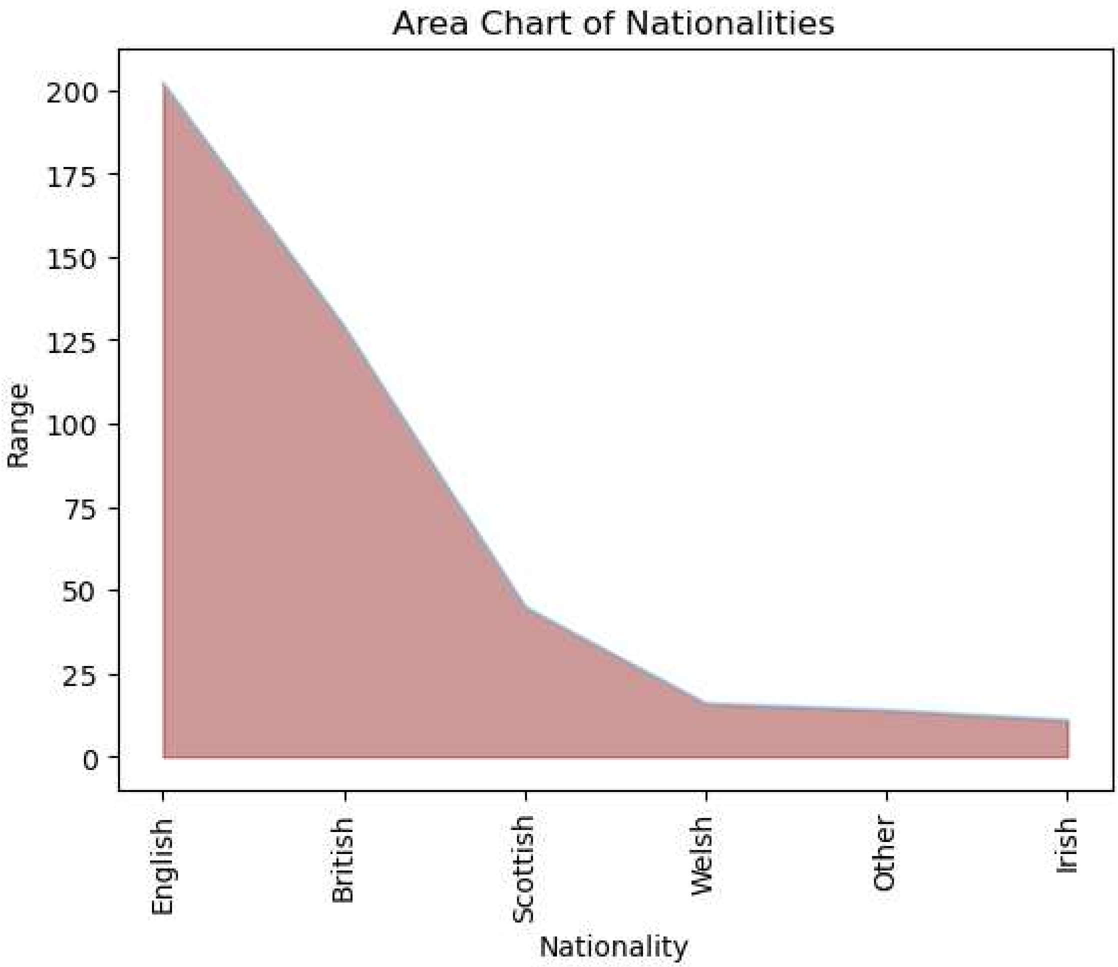
```
Out[40]: English     202
         British     129
         Scottish     45
         Welsh        16
         Other        14
         Irish        11
         Refused       3
         Unknown       1
         Name: Nationality, dtype: int64
```

```
In [41]: nationalities = ['English', 'British', 'Scottish', 'Welsh', 'Other', 'Irish']
         count = [202, 129, 45, 16, 14, 11]

         plt.fill_between(range(len(nationalities)), count, color='maroon', alpha=0.4)
         plt.plot(range(len(nationalities)), count, color='skyblue', alpha=0.6)

         plt.xticks(range(len(nationalities)), nationalities, rotation=90)
         plt.xlabel('Nationality')
         plt.ylabel('Range')
         plt.title('Area Chart of Nationalities')

         plt.show()
```



In [ ]: