Data analysis of smoking in UK citizens

Importing packages

In [48]:
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np
import seaborn as sns
```

In [3]:
```
df = pd.read_csv("smoking.csv")
```

Overview of the data

In [4]:
```
df.shape
```

Out[4]:
```
(1691, 12)
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1691 entries, 0 to 1690
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   gender                 1691 non-null   object 
 1   age                    1691 non-null   int64  
 2   marital_status         1691 non-null   object 
 3   highest_qualification  1691 non-null   object 
 4   nationality            1691 non-null   object 
 5   ethnicity              1691 non-null   object 
 6   gross_income           1691 non-null   object 
 7   region                 1691 non-null   object 
 8   smoke                  1691 non-null   object 
 9   amt_weekends           421 non-null    float64
 10  amt_weekdays           421 non-null    float64
 11  type                   421 non-null    object 
dtypes: float64(2), int64(1), object(9)
memory usage: 158.7+ KB
```

In [6]:
```
df.describe()
```

Out[6]:

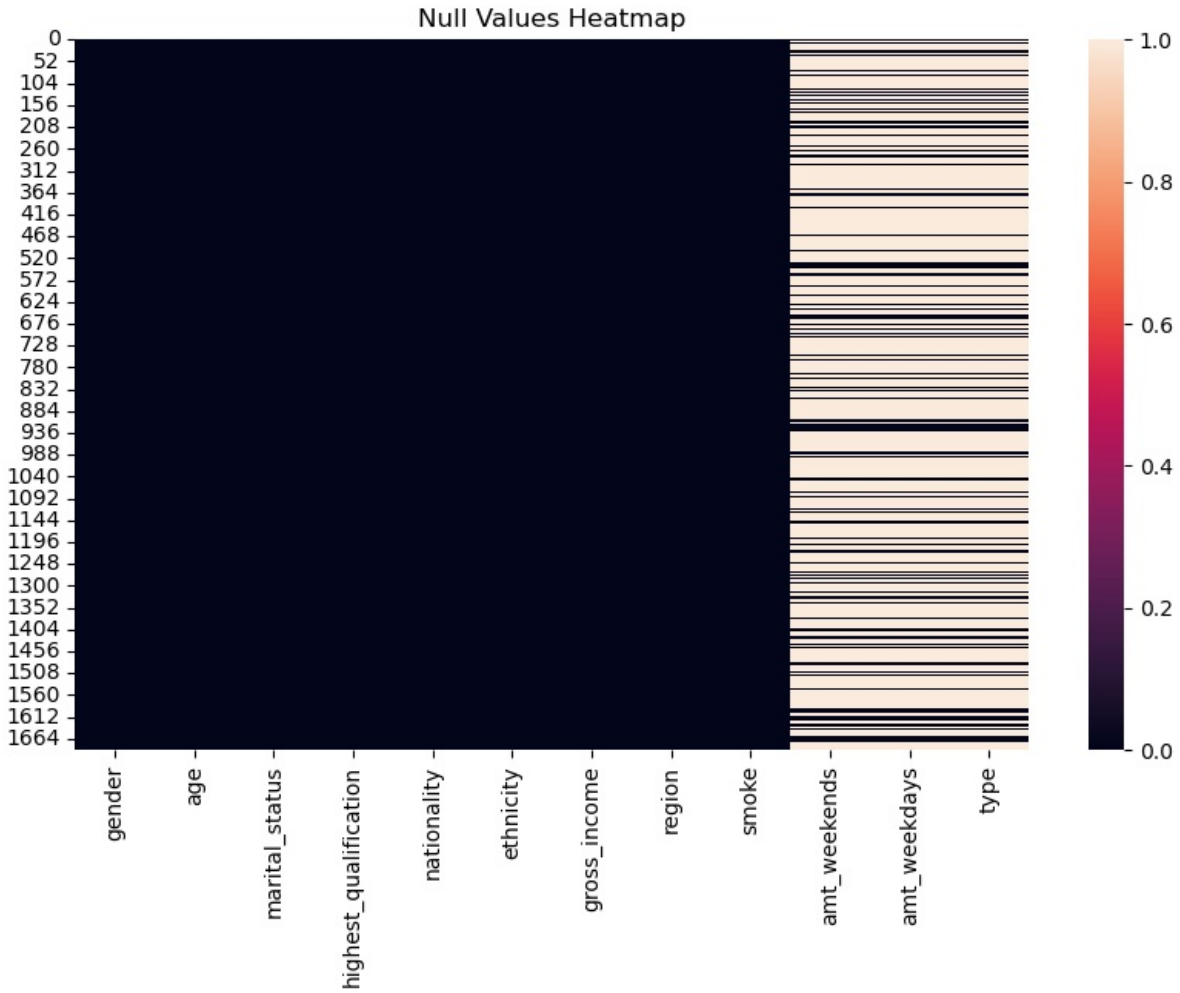
	age	amt_weekends	amt_weekdays
count	1691.000000	421.000000	421.000000
mean	49.836192	16.410926	13.750594
std	18.736851	9.892988	9.388292
min	16.000000	0.000000	0.000000
25%	34.000000	10.000000	7.000000
50%	48.000000	15.000000	12.000000
75%	65.500000	20.000000	20.000000
max	97.000000	60.000000	55.000000

In [7]:
```
df.isnull().value_counts()
```

Out[7]:
```
gender  age    marital_status  highest_qualification  nationality  ethnicity  gross_income  region  smoke  amt_
weekends  amt_weekdays  type
False   False  False           False                  False        False      False         False   False  True
True          True     1270

False         False         False     421
Name: count, dtype: int64
```

In [8]:
```
plt.figure(figsize=(10, 6))
sns.heatmap(df.isnull())
plt.title('Null Values Heatmap')
plt.show()
```



There seems to be empty values in 3 columns which will be dealth with later

In [9]:
```
df.select_dtypes(include=['float64', 'int64']).corr()
```

Out[9]:

	age	amt_weekends	amt_weekdays
age	1.000000	0.058642	0.192783
amt_weekends	0.058642	1.000000	0.802052
amt_weekdays	0.192783	0.802052	1.000000

Number of males vs females in the dataset

In [10]:
```
male = df[df["gender"] == "Male"]
female = df[df["gender"] == "Female"]

print(f"Number of males: {male.shape[0]}")
print(f"Number of females: {female.shape[0]}")
```

```
Number of males: 726
Number of females: 965
```

Numbers of male and females who smoke vs who don't

In [11]:
```
ms = male[male["smoke"] == "Yes"]
fs = female[female["smoke"] == "Yes"]

print(f"Number of males who some: {ms.shape[0]}, number who don't smoke: {male.shape[0] - ms.shape[0]}")
print(f"Number of females who smoke: {fs.shape[0]}, number who don't smoke: {female.shape[0] - fs.shape[0]}")
```

```
Number of males who some: 187, number who don't smoke: 539
Number of females who smoke: 234, number who don't smoke: 731
```

In [12]:
```
df_smoker = df[df["smoke"] == "Yes"]
```

How many cigarettes do smokers smoke daily on weekdays vs weekends

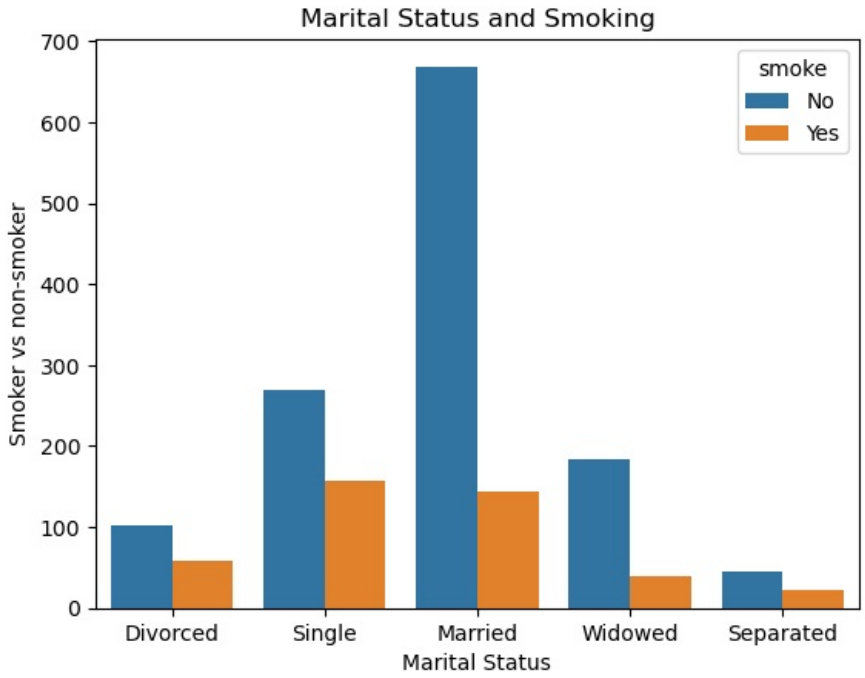
In [13]:
```
weekend = df_smoker["amt_weekends"].mean()
weekday = df_smoker["amt_weekdays"].mean()
```

In [14]:
```
print(f"Mean number of smoked cigarrettes based on weekend vs weekday\nWeekend: {weekend:.1f}\nWeekday: {weekda
```

```
Mean number of smoked cigarrettes based on weekend vs weekday
Weekend: 16.4
Weekday: 13.8
```

number of people who are smokers vs non-smokers based on their marital status

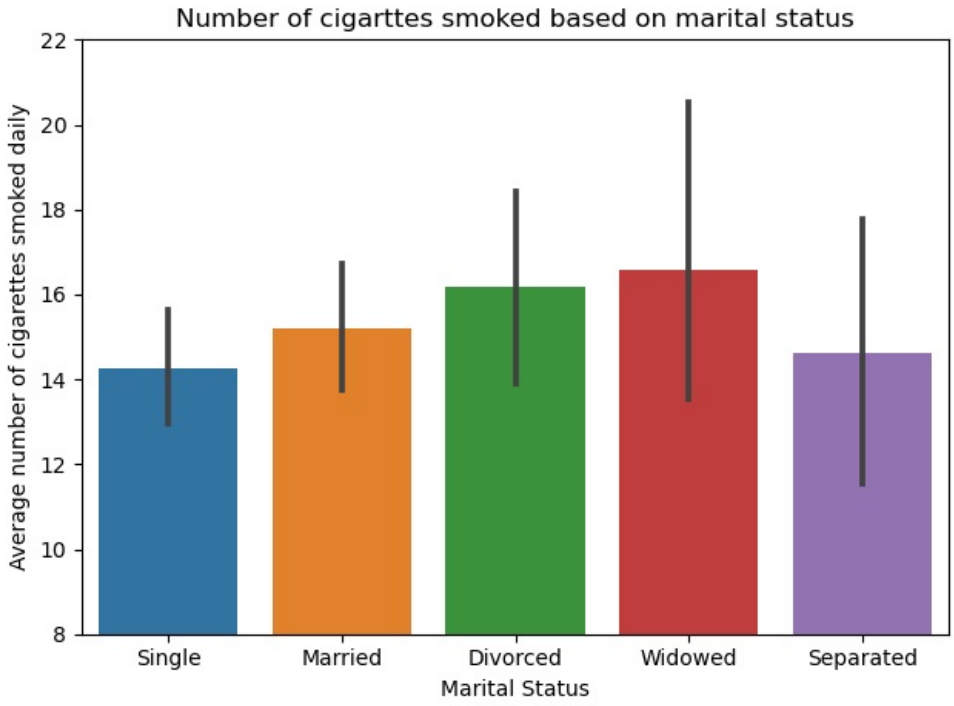
In [15]:
```
sns.countplot(x=df["marital_status"], hue=df["smoke"])
plt.xlabel("Marital Status")
plt.ylabel("Smoker vs non-smoker")
plt.title("Marital Status and Smoking")
plt.show()
```



Smoker rates seems to be much higher amongst single and divorced people, whereas it seems to be the lowest amongst married individuals

Here is a bar plot visualizing the amount of cigarettes smokers smoke based on their marital status

In [16]:
```
# Number of cigarettes smoked is expressed as the daily mean, both weekends and weekdays combined
sns.barplot(x=df_smoker["marital_status"], y=((df["amt_weekends"] + df["amt_weekdays"])/2))
plt.xlabel("Marital Status")
plt.ylabel("Average number of cigarettes smoked daily")
plt.ylim(8, 22)
plt.title("Number of cigarttes smoked based on marital status")
plt.tight_layout()
plt.show()
```

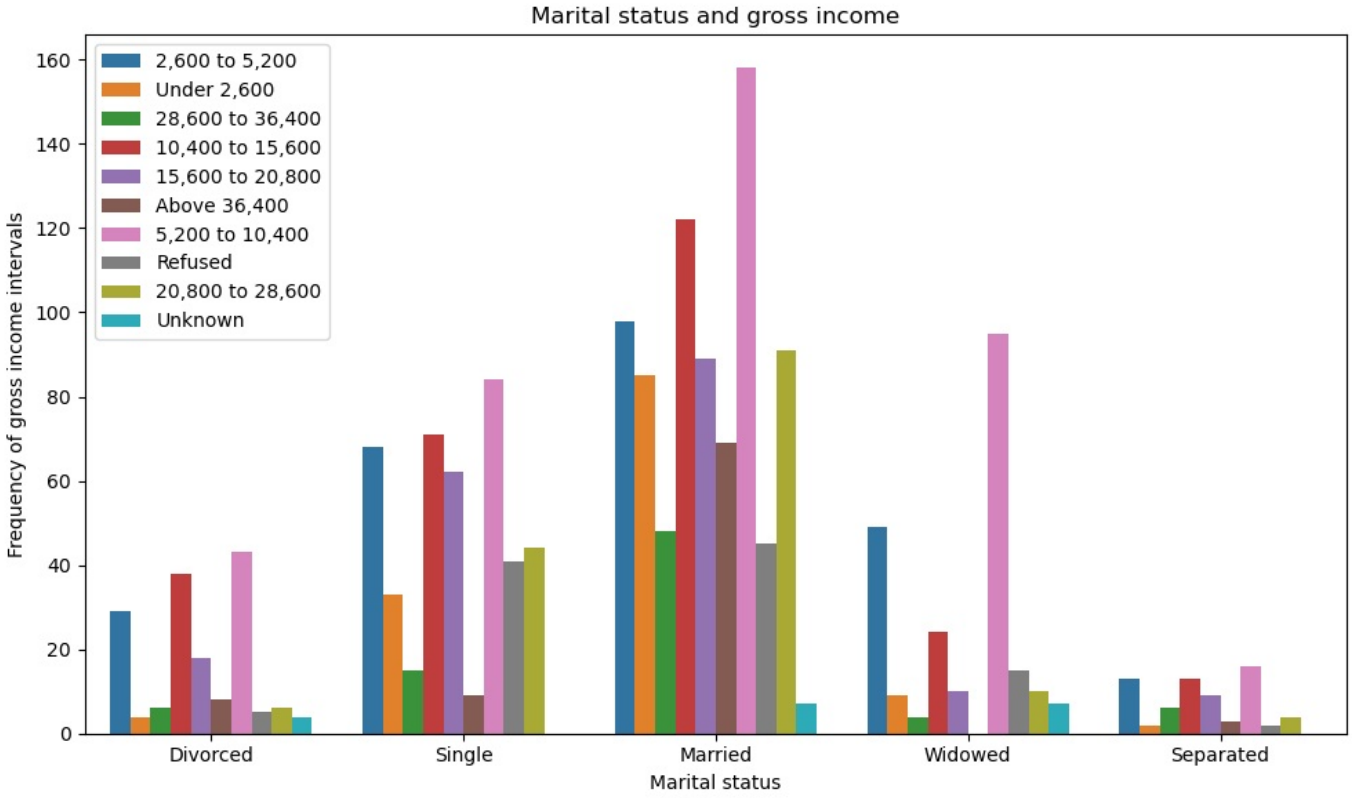


Although the difference isn't huge, divorced and widowed smokers smoke the most

Count plot visualizing marital status and gross income

In [17]:
```
plt.figure(figsize=(10, 6))
sns.countplot(x=df["marital_status"], hue=df["gross_income"])

plt.legend(loc='upper left')
plt.title("Marital status and gross income")
plt.xlabel("Marital status")
plt.ylabel("Frequency of gross income intervals")
plt.tight_layout()
```

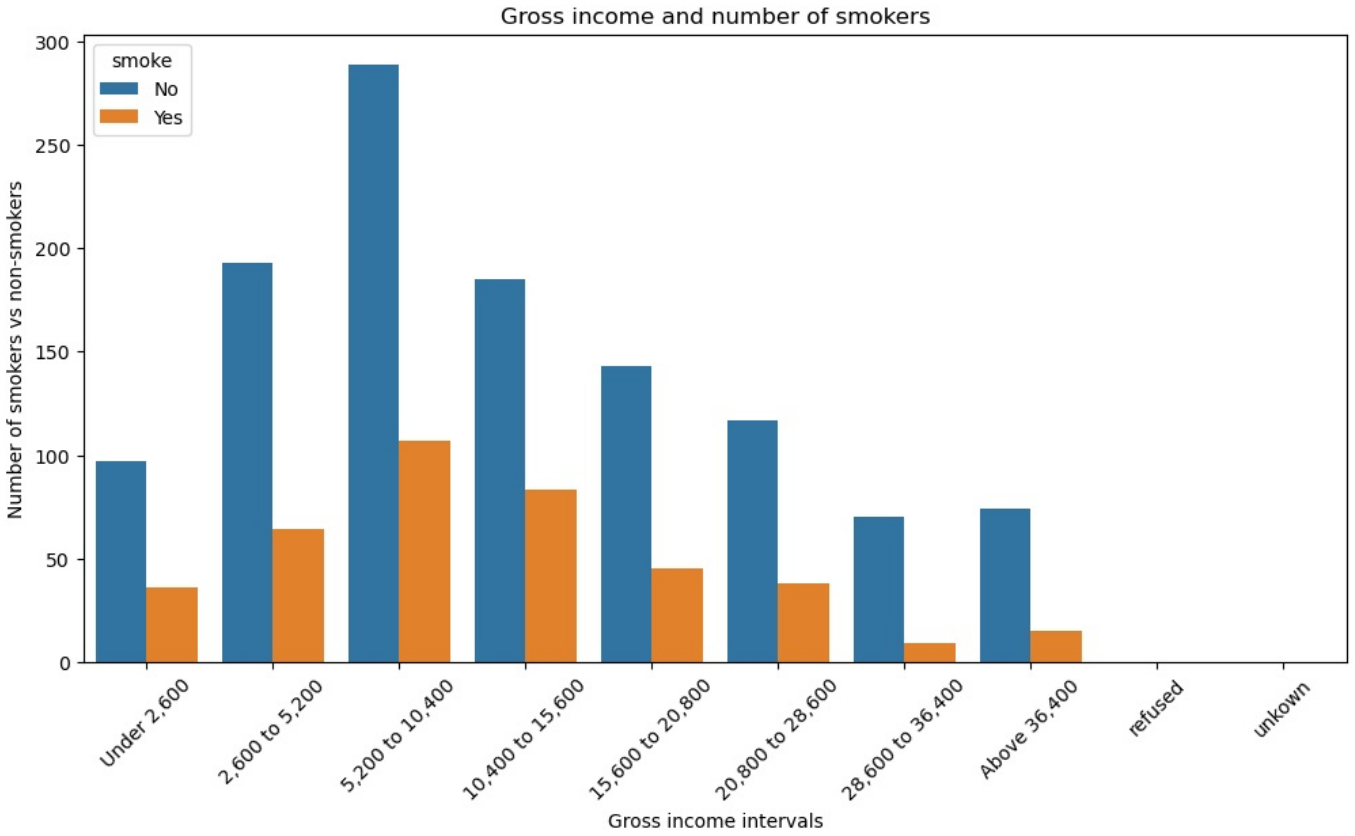


Married and single individuals seems to have the highest number of gross income rate falling into the category of "28,600 to 36,400" as well as higher rates of income in lower but high intervals

Now i will analyze the relation between smoking and gross income directly

In [18]:
```
order = ["Under 2,600", "2,600 to 5,200", "5,200 to 10,400", "10,400 to 15,600", "15,600 to 20,800", "20,800 to
# order the x axis respecctively
```

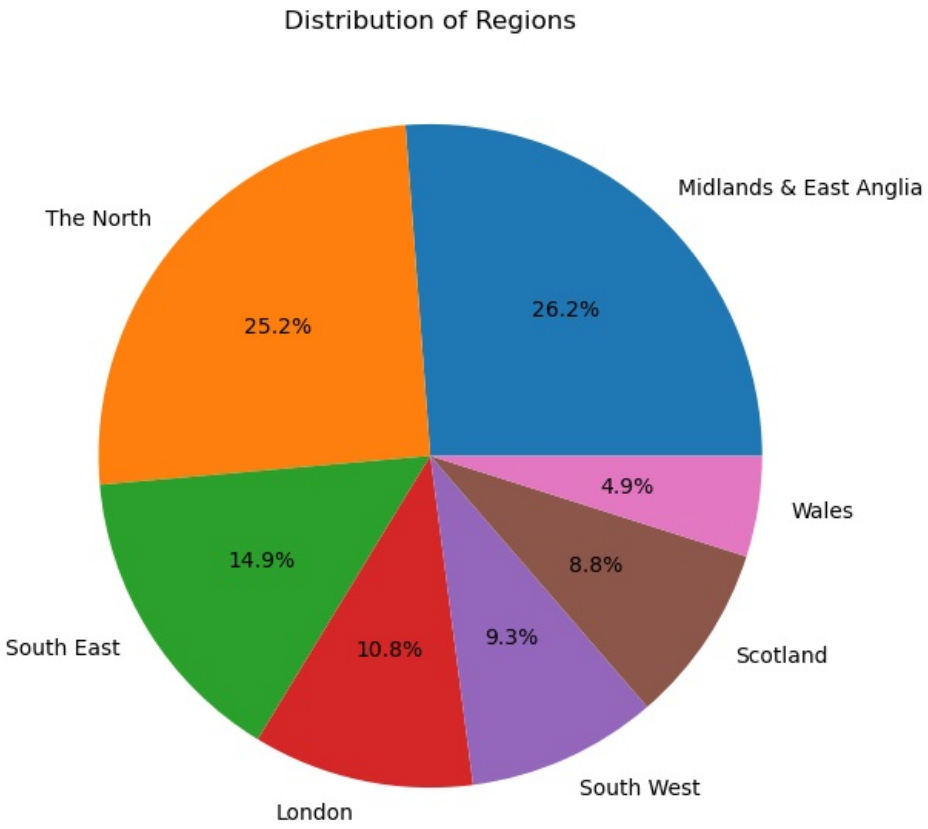
```
plt.figure(figsize=(12, 6))
sns.countplot(x=df["gross_income"], hue=df["smoke"], order=order)
plt.xticks(rotation=45)
plt.xlabel("Gross income intervals")
plt.ylabel("Number of smokers vs non-smokers")
plt.title("Gross income and number of smokers")
plt.show()
```



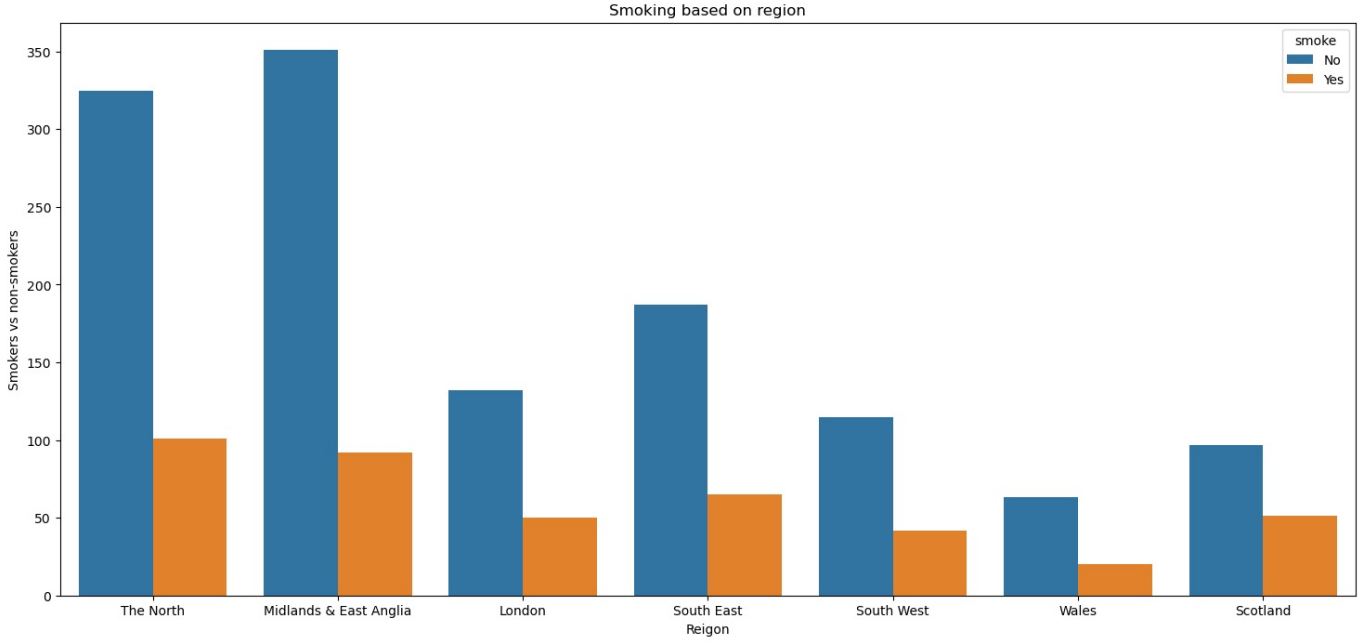
There seems to be a noticable relation between gross income and smoking but it is not a strong indicator.

Region based data analysis

In [19]:
```
region = df['region']
region_counts = region.value_counts()
plt.figure(figsize=(10, 7))
plt.pie(region_counts, labels=region_counts.index, autopct='%1.1f%%')
plt.title('Distribution of Regions')
plt.show()
```



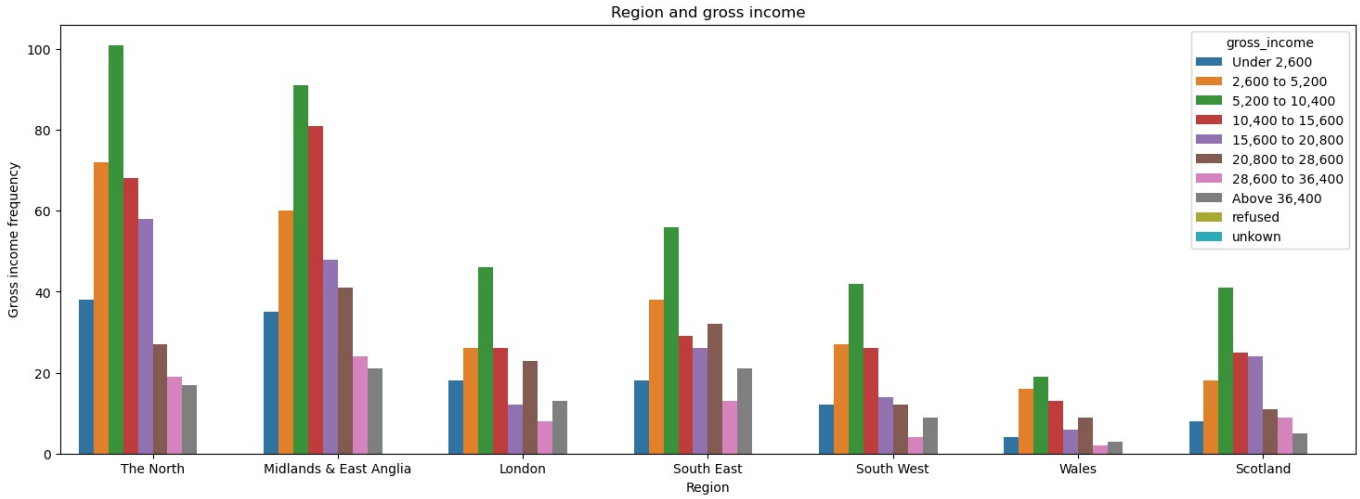
In [20]:
```
plt.figure(figsize=(18, 8))
sns.countplot(x=df["region"], hue=df["smoke"])
plt.xlabel("Reigon")
plt.ylabel("Smokers vs non-smokers")
plt.title("Smoking based on region")
plt.show()
```



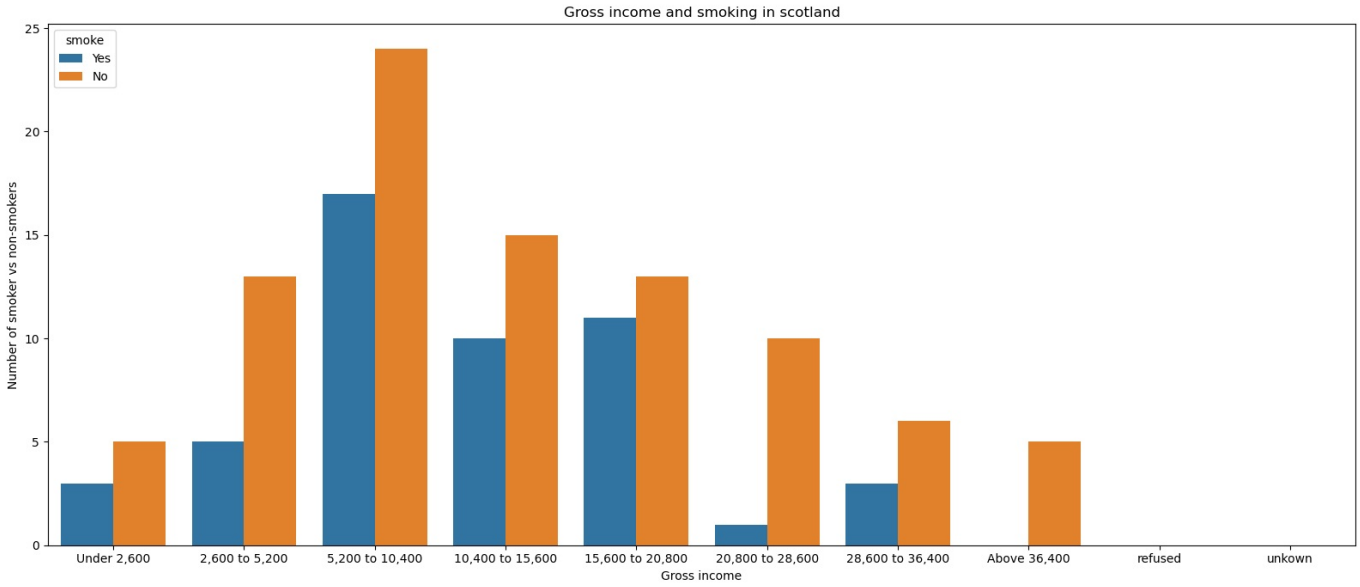
Smoking rates seems to be the highest in scotland

Gross income and region data alanysis

In [43]:
```
plt.figure(figsize=(18,6))
sns.countplot(x=df["region"], hue=df["gross_income"], hue_order=order)
plt.xlabel("Region")
plt.ylabel("Gross income frequency")
plt.title("Region and gross income")
plt.show()
```



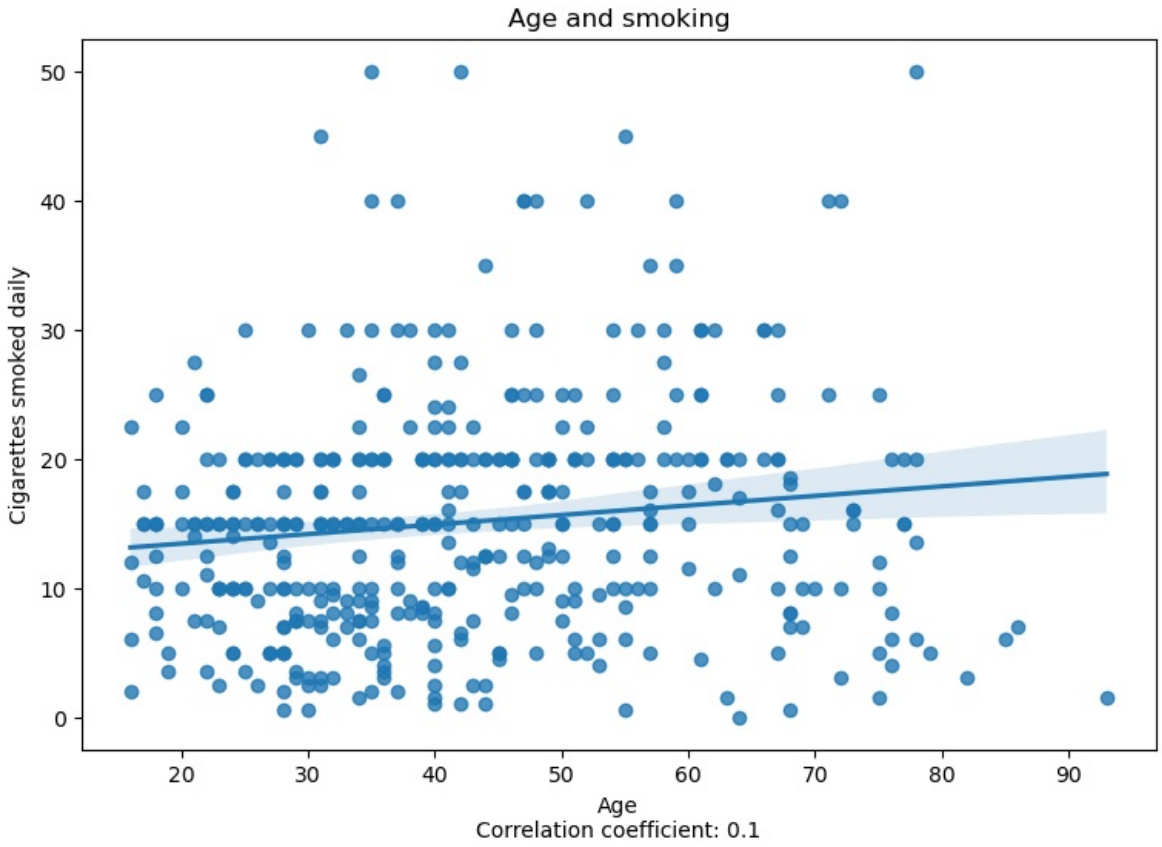
In [31]:
```
scotland_data = df[df["region"] == "Scotland"]
plt.figure(figsize=(16,7))
sns.countplot(x=scotland_data["gross_income"], hue=scotland_data["smoke"], order=order)
plt.xlabel("Gross income")
plt.ylabel("Number of smoker vs non-smokers")
plt.title("Gross income and smoking in scotland")
plt.tight_layout()
plt.show()
```



People in scotland with higher levels of income seem to be less likely to smoke overall

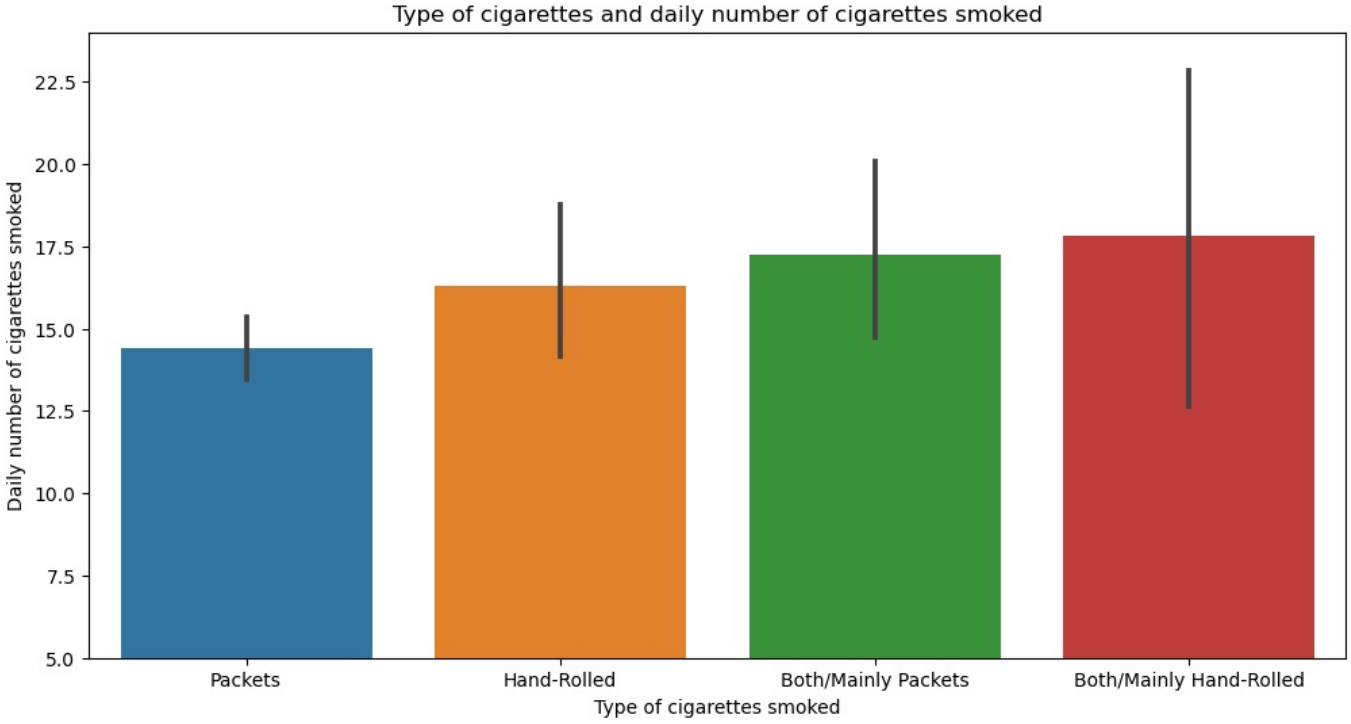
Smoking and Age

In [64]:
```
age1 = df_smoker["age"]
smoke1 = ((df_smoker["amt_weekdays"] + df_smoker["amt_weekends"])/2)
plt.figure(figsize=(9,6))
sns.regplot(x=age1, y=smoke1)
plt.xlabel(f"Age\nCorrelation coefficient: {age1.corr(smoke1):.1f}")
plt.ylabel("Cigarettes smoked daily")
plt.title("Age and smoking")
plt.show()
```



Type of cigarettes

In [70]:
```
plt.figure(figsize=(12,6))
sns.barplot(x="type", y=smoke1, data=df_smoker)
plt.xlabel("Type of cigarettes smoked")
plt.ylabel("Daily number of cigarettes smoked")
plt.title("Type of cigarettes and daily number of cigarettes smoked")
plt.ylim(5,24)
plt.show()
```



In [74]:
```
plt.figure(figsize=(12,6))
sns.countplot(x=df_smoker["region"], hue=df_smoker["type"])
plt.title("Region and type of cigarettes smoked")
plt.xlabel("Region")
plt.ylabel("Frequency of each cigarette type")
plt.title("Region and type of cigarettes")
plt.show()
```