Miles A. Peña

DSC 550

Week 2 Exercise

03/24/2024

Using a data set of your choice, write an introduction explaining the data set.

The dataset presented aims to determine any relationship that might exist between smoking and other factors such as gender and ethnicity. Its purpose is to analyze patterns between the variables aforementioned as well as others like age, marital status, education, and nationality to name a few.

Identify a question or question(s) that you would like to explore in your data set.

1) Does marital status affect whether or not a person smokes? 2) Are female individuals or male individuals more likely to be smokers? 3) How does age affect if an individual smokes or not?

Create at least three graphs that help answer these questions. Make sure your graphs are clearly readable and are labeled appropriately and professionally.

In [54]:
```
# download necessary packages

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

import matplotlib as mpl
```

In [55]:
```
# import the dataset

smoking = pd.read_csv('smoking.csv')
smoking
```

Out[55]:

	Unnamed: 0	gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_income	region	smoke	amt_weekends	amt_weekdays	type
0	1	Male	38	Divorced	No Qualification	British	White	2,600 to 5,200	The North	No	NaN	NaN	NaN
1	2	Female	42	Single	No Qualification	British	White	Under 2,600	The North	Yes	12.0	12.0	Packets
2	3	Male	40	Married	Degree	English	White	28,600 to 36,400	The North	No	NaN	NaN	NaN
3	4	Female	40	Married	Degree	English	White	10,400 to 15,600	The North	No	NaN	NaN	NaN
4	5	Female	39	Married	GCSE/O Level	British	White	2,600 to 5,200	The North	No	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1686	1687	Male	22	Single	No Qualification	Scottish	White	2,600 to 5,200	Scotland	No	NaN	NaN	NaN
1687	1688	Female	49	Divorced	Other/Sub Degree	English	White	2,600 to 5,200	Scotland	Yes	20.0	20.0	Hand-Rolled
1688	1689	Male	45	Married	Other/Sub Degree	Scottish	White	5,200 to 10,400	Scotland	No	NaN	NaN	NaN
1689	1690	Female	51	Married	No Qualification	English	White	2,600 to 5,200	Scotland	Yes	20.0	20.0	Packets
1690	1691	Male	31	Married	Degree	Scottish	White	10,400 to 15,600	Scotland	No	NaN	NaN	NaN

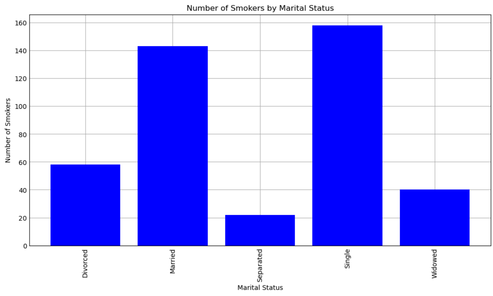
1691 rows × 13 columns

In [75]:
```
# determine the number of smokers for each of the 5 marital statuses

smoker_marital_yes = smoking[smoking['smoke'] == 'Yes'].groupby('marital_status').size()
smoker_marital_yes
```

Out[75]:
```
marital_status
Divorced     58
Married     143
Separated    22
Single      158
Widowed      40
dtype: int64
```

In [128]:
```
plt.figure(figsize = (10, 5))
plt.bar(smoker_marital_yes.index, smoker_marital_yes.values, color = 'blue')
plt.xlabel('Marital Status')
plt.ylabel('Number of Smokers')
plt.title('Number of Smokers by Marital Status')
# rotate x-axis labels since they do not fill well horizontally
plt.xticks(smoker_marital_yes.index, rotation = 90)
plt.show()
```

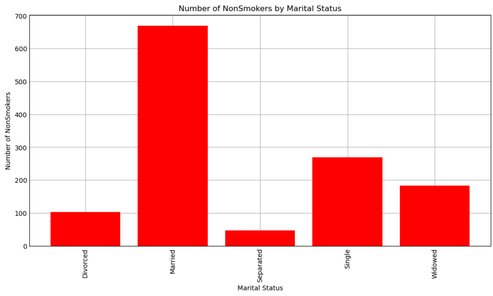


In [96]:
```
# determine the number of non-smokers for each of the 5 marital statuses

smoker_marital_no = smoking[smoking['smoke'] == 'No'].groupby('marital_status').size()
smoker_marital_no
```

Out[96]:
```
marital_status
Divorced     103
Married      669
Separated     46
Single       269
Widowed      183
dtype: int64
```

In [127]:
```
plt.figure(figsize = (10, 5))
plt.bar(smoker_marital_no.index, smoker_marital_no.values, color = 'red')
plt.xlabel('Marital Status')
plt.ylabel('Number of NonSmokers')
plt.title('Number of NonSmokers by Marital Status')
# rotate x-axis labels since they do not fill well horizontally
plt.xticks(smoker_marital_no.index, rotation = 90)
plt.show()
```

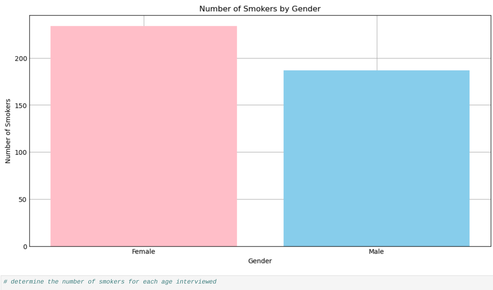


In [99]:
```
# determine number of smokers for each of the two available gender markers

smoker_by_gender = smoking[smoking['smoke'] == 'Yes'].groupby('gender').size()
smoker_by_gender
```

Out[99]:
```
gender
Female    234
Male      187
dtype: int64
```

In [129]:
```
plt.figure(figsize = (10, 5))
# change the color by bar in order as they appear in dataset
plt.bar(smoker_by_gender.index, smoker_by_gender.values, color = ['pink', 'skyblue'])
plt.xlabel('Gender')
plt.ylabel('Number of Smokers')
plt.title('Number of Smokers by Gender')
plt.show()
```

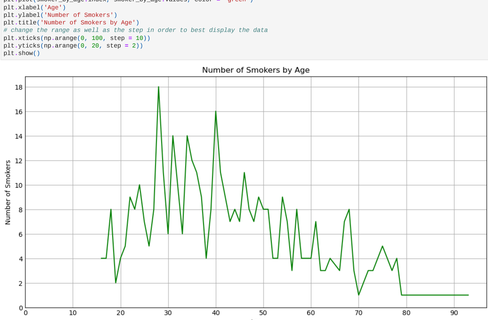


In [117]:
```
# determine the number of smokers for each age interviewed

smoker_by_age = smoking[smoking['smoke'] == 'Yes'].groupby('age').size()
smoker_by_age
```

Out[117]:
```
age
16    4
17    4
18    8
19    2
20    4
     ..
79    1
82    1
85    1
86    1
93    1
Length: 66, dtype: int64
```

In [130]:
```
plt.figure(figsize = (10, 5))
plt.plot(smoker_by_age.index, smoker_by_age.values, color = 'green')
plt.xlabel('Age')
plt.ylabel('Number of Smokers')
plt.title('Number of Smokers by Age')
# change the range as well as the step in order to best display the data
plt.xticks(np.arange(0, 100, step = 10))
plt.yticks(np.arange(0, 20, step = 2))
plt.show()
```



Explain what you have learned from each of your graphs.

Graphs 1 and **2** show that when looking at the data analyzing whether or not marital status affects if a person smokes or not we can see off the bat that the number of nonsmokers in general is greater than those who do smoke. Due to this change of the y-axis, it would be easy to get confused and get impacted by the images. However, after looking further into the data we can see that marital status does not seem to have too much of an effect on whether or not individuals pick up smoking. For all marital statuses, there are more nonsmokers than there are smokers and thus we cannot confidently determine a relationship.

In **Graph 3** we see that the analysis of smokers by gender helps us get a clearer picture than the last method. Based on the visualization, we can see that according to the data set females are in fact more likely to be smokers than their male counterparts with females being 47 individuals higher in the total count.

By looking at the data of smokers by age in **Graph 4**, we can see that the majority of smokers are between the ages of about 28 and 40. However, there are some fluctuations even within this range that prevent us from making a determination as to whether or not age is a factor that influences if an individual smokes or not. What we can tell from the data however, is that individuals of age 28 are the most accounted for in the dataset as smokers and that as the individual gets older, we can see the trend reduce in numbers meaning that as time passes people lay off the smoking. At about the age of 45 we can start to see this decline happen.

Write a conclusion that summarizes your findings.

After conducting the analysis on the dataset, it does not appear that marital status has an influence on smoking behavior. The examination of the dataset revealed varying numbers of smokers across different marital statuses, with no actual trend. When it comes to analyzing the data in regards to gender, it's evident that gender plays a role in determining smoking behavior. The data revealed that there are 187 male smokers and 234 female smokers for this particular set. However, we must make it clear that further analysis may be needed to explore potential factors influencing this disparity, such as social norms, cultural differences, or individual preferences. It also allows for the traditional assumption that men are the more likely to smoke to be discarded and can assist with considering gender-specific approaches in public health initiatives aimed at tobacco control and smoking cessation efforts. In addition, there is variability in smoking prevalence across different age groups. There does not appear to be a direct correlation but the trend that can be seen is that as individuals get older, they tend to be less likely to smoke. Due to the fact that there are multiple variables that can account to whether an individual smokes or not, it is necessary to consider an approach that takes into account multiple determinants of smoking behavior in order to develop comprehensive tobacco control strategies and encourage healthier lifestyles across all individuals.