In [ ]:
```
#'Research Project #3: Continuation on the Comparison of Health Factors
##SDS 348 Spring 2021
```

In [ ]:
```
###Channing Burba, cb48596
```

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

In [2]:
```
#Import dataset 'PulseP2'
health = pd.read_csv("PulseP2.csv", index_col = 0)
print(health)
```

```
     Active  Rest  Smoke     Sex  Exercise  Hgt  Wgt
1        97    78      0  female         1   63  119
2        82    68      1    male         3   70  225
3        88    62      0    male         3   72  175
4       106    74      0    male         3   72  170
5        78    63      0  female         3   67  125
..      ...   ...    ...     ...       ...  ...  ...
228     105    85      0  female         2   64  150
229      82    74      0  female         3   66  124
230     102    81      0    male         2   69  172
231      87    67      0    male         2   68  170
232      81    62      0    male         3   68  151

[232 rows x 7 columns]
```

In [4]:
```
#Returns information about 'PulseP2' dataset
health.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 232 entries, 1 to 232
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Active    232 non-null    int64 
 1   Rest      232 non-null    int64 
 2   Smoke     232 non-null    int64 
 3   Sex       232 non-null    object
 4   Exercise  232 non-null    int64 
 5   Hgt       232 non-null    int64 
 6   Wgt       232 non-null    int64 
dtypes: int64(6), object(1)
memory usage: 14.5+ KB
```

The dataset that I chose, 'PulseP2,' measures the pulse rate of individuals prior to and after exercise. This dataset had 232 observations and 7 variables. The variables include: 'Active'-The pulse rate of an individual prior to and after exercise, 'Rest'-The participants regular resting pulse, 'Smoke'-An indication on whether the participant is a smoker or non-smoker, 'Sex'-The gender of the individual, 'Exercise'-The number of hours the participant exercises per week, 'Hgt'-The individual's height, and 'Wgt'-The individual's weight. The 'PulseP2' dataset was acquired by both survey and experimentation. The participant's pulses were recorded before and after walking 3 laps up and down a set of stairs, while information regarding their height, weight, sex, and smoking status was acquired through survey.

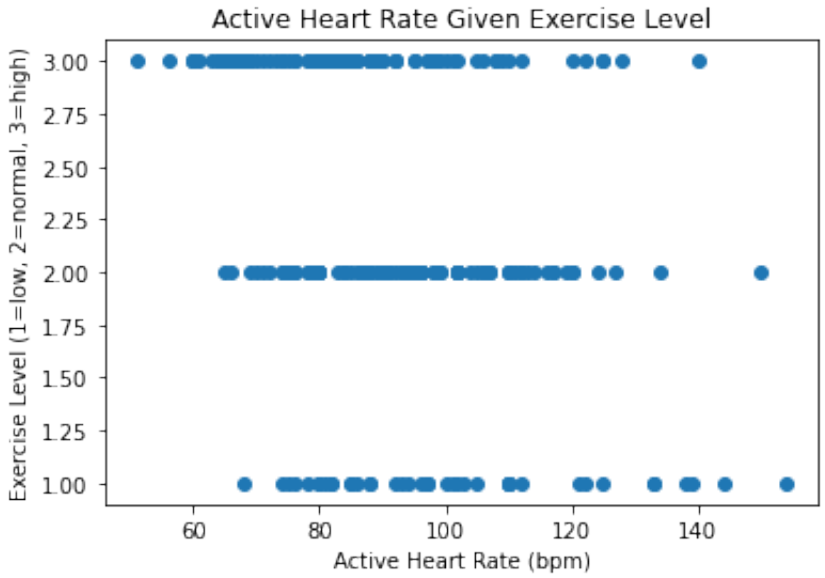
In [5]:
```
#Returns numerical statistics about 'PulseP2' dataset
health.describe()
```

Out[5]:

	Active	Rest	Smoke	Exercise	Hgt	Wgt
count	232.000000	232.000000	232.000000	232.000000	232.000000	232.000000
mean	91.297414	68.349138	0.112069	2.254310	68.245690	157.918103
std	18.820234	9.949378	0.316133	0.738536	3.738761	31.832587
min	51.000000	43.000000	0.000000	1.000000	60.000000	102.000000
25%	79.000000	62.000000	0.000000	2.000000	65.000000	135.000000
50%	88.500000	68.000000	0.000000	2.000000	68.000000	150.000000
75%	102.000000	74.000000	0.000000	3.000000	71.000000	175.000000
max	154.000000	106.000000	1.000000	3.000000	78.000000	260.000000

For the 'Active' heart rate of participants, the maximun heart rate is 154 bpm. This indicates that out of every participant, one participant had the highest heart rate of 154 bpm. This is also an indicator that this particular participant's heart has to work harder to supply blood to the rest of the body due to factors such as an unhealthy lifestyle and being a smoker. On the contrary, the minimum heart rate is 51 bpm. This indicates that out of every participant, one participant had the lowest heart rate of 51 bpm. This is also an indicator that this particular participant's heart has doesn't need to work very hard to supply blood to the rest of the body due to their healthy lifestyle and activity/exercise level. The standard deviation of the 'Active' heart rate is 18.820234 bpm. The standard deviation measures the dispersion of values. In this experiment, the mean 'Active' heart rate is 91.297414 bpm, so the heart rates are dispersed as either 18.820234 bpm above or below the mean bpm.

In [8]:
```
#Scatterplot: relationship between 'Active' heart rate and 'Exercise'
plt.scatter(health.Active, health.Exercise)
plt.title("Active Heart Rate Given Exercise Level")
plt.xlabel("Active Heart Rate (bpm)")
plt.ylabel("Exercise Level (1=low, 2=normal, 3=high)")
plt.show()
```



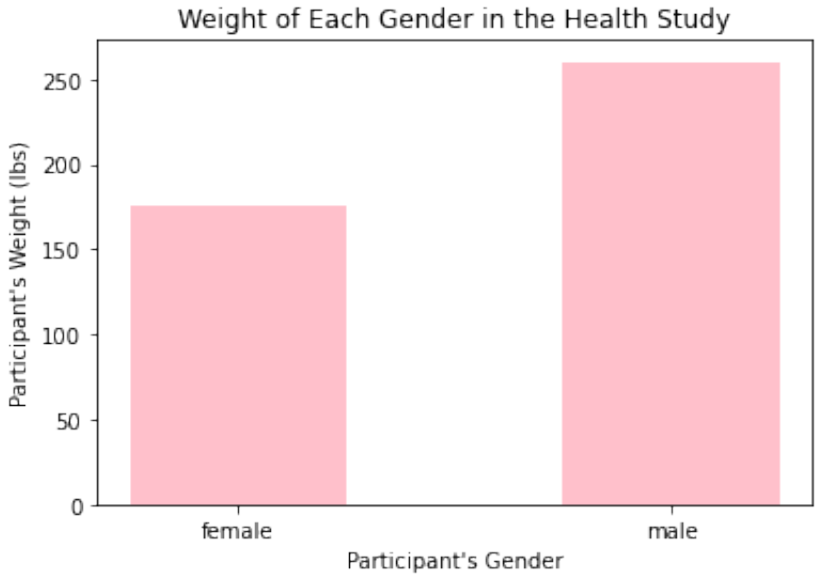
When looking at the scatterplot produced, it is clear to see that participants that have a low 'Active' heart rate also have a high 'Exercise' level. When excluding outliers, I came to the conclusion that as the 'Active' heart rate inceases, there is an inverse relationship with 'Exercise' level (i.e. there is a relationship between low 'Exercise' level and high 'Active' heart rate). The plot also shows that individuals are very consistant when it comes to the normality of each variable, as participants who 'Exercise' a normal amount tend to also have a normal 'Active' heart rate.

In [10]:
```
#Decribes 'Sex' categorical variable
health['Sex'].value_counts()
```

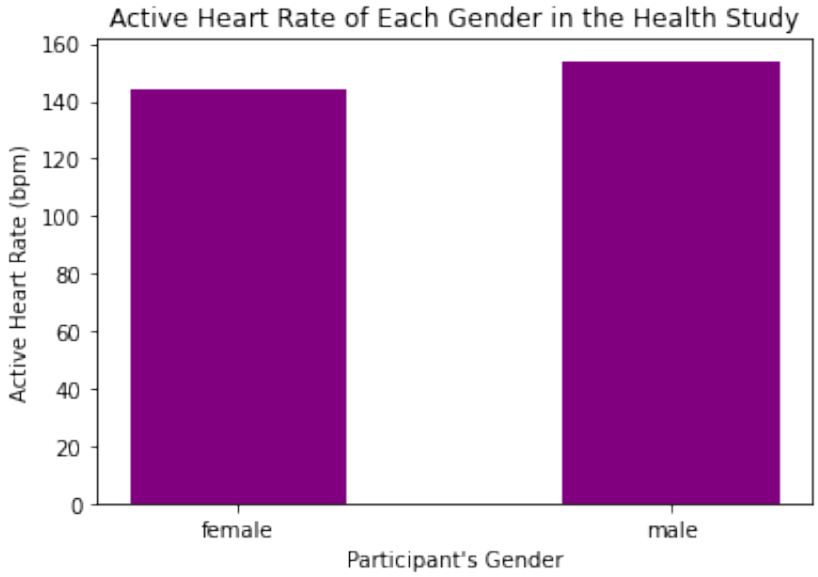
Out[10]:
```
male      122
female    110
Name: Sex, dtype: int64
```

When it comes to categorical data, value counts are the only way to statistically describe the data. For this experiment, 'Sex' was the categorical variable. The counts from above show the genders of each participant in the study. In this experiment, males were the group with the largest count, with a total of 122 males. Females encompassed the smallest group, having only 110 females in the study. In total, there were 232 individuals that participated in the study.

In [22]:
```
#Barplot: shows the relationship of 'Sex' and 'Wgt'
plt.bar(health.Sex, health.Wgt, color ='pink', width = 0.5)
plt.xlabel("Participant's Gender")
plt.ylabel("Participant's Weight (lbs)")
plt.title("Weight of Each Gender in the Health Study")
plt.show()
```



In [21]:
```
#Barplot: shows the relationship of 'Sex' and 'Active' heart rate
plt.bar(health.Sex, health.Active, color ='purple', width = 0.5)
plt.xlabel("Participant's Gender")
plt.ylabel("Active Heart Rate (bpm)")
plt.title("Active Heart Rate of Each Gender in the Health Study")
plt.show()
```



The two barplots above signify the differences in both 'Wgt'of participamts and the'Active' heart rates of males and females in the study. In the first plot, I wanted to see the difference in weight for each gender in order to see if weight in turn had an affect on the 'Active' heart rate of participants. For barplot 1, the women in the study weighed around 180 pounds, whereas the men in the study weighed about 250 pounds. In the second barplot, men have a slightly higher 'Active' heart rate than females (i.e. 140 bpm for females and 160 bpm for males). This makes sense because the men in the study have a higher body weight, therefore influencing a higher 'Active' heart rate due to their body having to produce a higher blood supply to support a bigger body. Another reason there is a slight difference in 'Active' heart rate is due to some participants (men and women) being smokers, as smoking can increase 'Active' heart rate.