

```
In [166]: print("Name : ")
print("We will be cleaning the big data and make a comparison to show who has a h
print("Also we will be deriviring which age group has the high chances of coronar
```

Name :
 We will be cleaning the big data and make a comparison to show who has a health
 ier heart smokers OR non smokers, usings a line graph
 Also we will be deriviring which age group has the high chances of coronary hea
 rt disease in 10 years

Task 1 - Plot a line graph to show the difference between heart rate of smokers and non smokers

```
In [156]: #Import Libraries
import pandas as pd
import matplotlib.pyplot as plt

#read the csv
df = pd.read_csv('framingham.csv')
df
```

```
Out[156]:
```

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | c |
|------|------|-----|-----------|---------------|------------|--------|-----------------|--------------|-----|
| 0 | 1 | 39 | 4.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 1 | 0 | 46 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 2 | 1 | 48 | 1.0 | 1 | 20.0 | 0.0 | 0 | 0 | |
| 3 | 0 | 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | 1 | |
| 4 | 0 | 46 | 3.0 | 1 | 23.0 | 0.0 | 0 | 0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4233 | 1 | 50 | 1.0 | 1 | 1.0 | 0.0 | 0 | 1 | |
| 4234 | 1 | 51 | 3.0 | 1 | 43.0 | 0.0 | 0 | 0 | |
| 4235 | 0 | 48 | 2.0 | 1 | 20.0 | NaN | 0 | 0 | |
| 4236 | 0 | 44 | 1.0 | 1 | 15.0 | 0.0 | 0 | 0 | |
| 4237 | 0 | 52 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | |

4238 rows × 16 columns

In [157]: *#Filter and make a new dataframe for non smokers*

```
non_smokers = df.loc[df['currentSmoker'] == 0]

non_smokers
```

Out[157]:

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | c |
|------|------|-----|-----------|---------------|------------|--------|-----------------|--------------|-----|
| 0 | 1 | 39 | 4.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 1 | 0 | 46 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 5 | 0 | 43 | 2.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| 6 | 0 | 63 | 1.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 8 | 1 | 52 | 1.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4226 | 1 | 58 | 1.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 4228 | 0 | 50 | 1.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| 4231 | 1 | 58 | 3.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| 4232 | 1 | 68 | 1.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| 4237 | 0 | 52 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | |

2144 rows × 16 columns



In [158]: *#Group by age column and find average heart rate at different age among the non s*

```
group_non_smokers = non_smokers.groupby('age')['heartRate'].mean().reset_index()

group_non_smokers
```

| | | |
|----|----|-----------|
| 25 | 58 | 75.343750 |
| 26 | 59 | 74.197368 |
| 27 | 60 | 75.342857 |
| 28 | 61 | 73.770270 |
| 29 | 62 | 74.202899 |
| 30 | 63 | 75.129870 |
| 31 | 64 | 76.469697 |
| 32 | 65 | 74.200000 |
| 33 | 66 | 80.714286 |
| 34 | 67 | 73.448276 |
| 35 | 68 | 80.166667 |
| 36 | 69 | 80.500000 |
| 37 | 70 | 64.000000 |



```
In [159]: #Filter and make a new dataframe for smokers
smokers = df.loc[df['currentSmoker'] == 1]
smokers
```

Out[159]:

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | c |
|------|------|-----|-----------|---------------|------------|--------|-----------------|--------------|-----|
| 2 | 1 | 48 | 1.0 | 1 | 20.0 | 0.0 | 0 | 0 | |
| 3 | 0 | 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | 1 | |
| 4 | 0 | 46 | 3.0 | 1 | 23.0 | 0.0 | 0 | 0 | |
| 7 | 0 | 45 | 2.0 | 1 | 20.0 | 0.0 | 0 | 0 | |
| 9 | 1 | 43 | 1.0 | 1 | 30.0 | 0.0 | 0 | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4230 | 0 | 56 | 1.0 | 1 | 3.0 | 0.0 | 0 | 1 | |
| 4233 | 1 | 50 | 1.0 | 1 | 1.0 | 0.0 | 0 | 1 | |
| 4234 | 1 | 51 | 3.0 | 1 | 43.0 | 0.0 | 0 | 0 | |
| 4235 | 0 | 48 | 2.0 | 1 | 20.0 | NaN | 0 | 0 | |
| 4236 | 0 | 44 | 1.0 | 1 | 15.0 | 0.0 | 0 | 0 | |

2094 rows × 16 columns



```
In [162]: #Group by age column and find average heart rate at different age among the smokers  
group_smokers = smokers.groupby('age')['heartRate'].mean().reset_index()  
group_smokers
```

Out[162]:

| | age | heartRate |
|----|-----|-----------|
| 0 | 32 | 80.000000 |
| 1 | 33 | 75.000000 |
| 2 | 34 | 73.272727 |
| 3 | 35 | 72.956522 |
| 4 | 36 | 75.191489 |
| 5 | 37 | 74.410714 |
| 6 | 38 | 79.431818 |
| 7 | 39 | 75.755102 |
| 8 | 40 | 75.727273 |
| 9 | 41 | 78.019608 |
| 10 | 42 | 75.827273 |
| 11 | 43 | 79.151515 |
| 12 | 44 | 75.553398 |
| 13 | 45 | 75.364583 |
| 14 | 46 | 78.752381 |
| 15 | 47 | 76.775281 |
| 16 | 48 | 76.852273 |
| 17 | 49 | 76.647887 |
| 18 | 50 | 77.068493 |
| 19 | 51 | 77.315789 |
| 20 | 52 | 75.775862 |
| 21 | 53 | 77.474576 |
| 22 | 54 | 77.384615 |
| 23 | 55 | 76.220000 |
| 24 | 56 | 74.527273 |
| 25 | 57 | 77.039216 |
| 26 | 58 | 78.283019 |
| 27 | 59 | 75.604651 |
| 28 | 60 | 77.390244 |
| 29 | 61 | 76.555556 |
| 30 | 62 | 75.033333 |
| 31 | 63 | 72.303030 |

| | age | heartRate |
|----|-----|-----------|
| 32 | 64 | 77.307692 |
| 33 | 65 | 75.500000 |
| 34 | 66 | 73.700000 |
| 35 | 67 | 81.375000 |
| 36 | 68 | 80.500000 |
| 37 | 69 | 72.333333 |

```
In [165]: #Plot a line graph to show the heart rate of smokers vs non smokers

plt.figure(figsize=(20, 8))
plt.plot(group_non_smokers['age'], group_non_smokers['heartRate'], label = "Non S

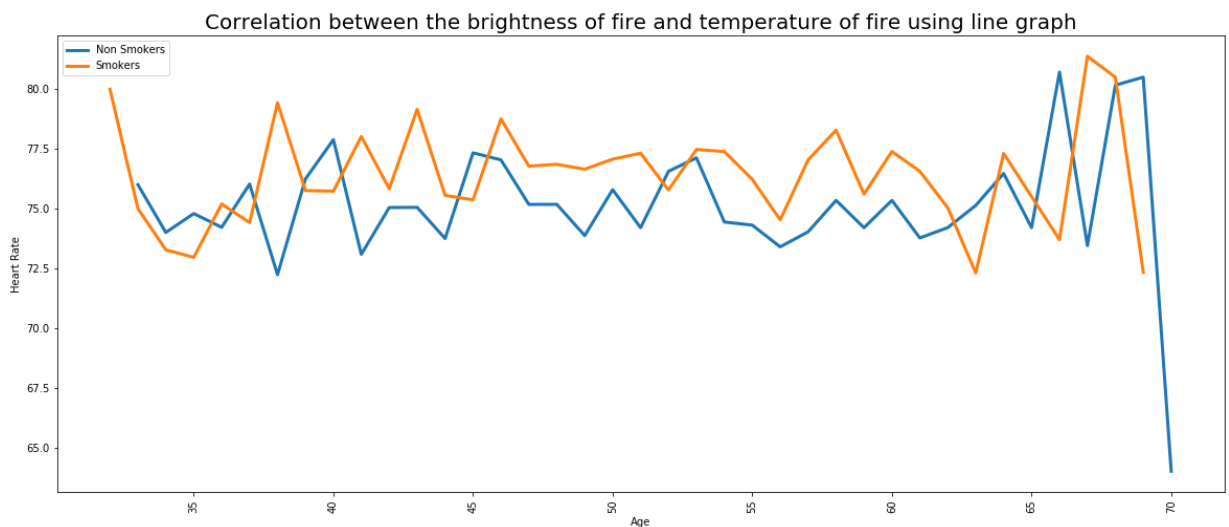
plt.plot(group_smokers['age'], group_smokers['heartRate'], label = "Smokers", lin

plt.xlabel('Age')
plt.ylabel('Heart Rate')
plt.xticks(rotation='vertical')

plt.title('Correlation between the heart rate of smokers and non smokers using li

plt.legend()

plt.show()
```



Conslusion - The heart rate of non smokers are lesser than the people who smoke, hence we conclude that the heart of the non smokers are healthier than that of smokers

Task 2 - Which age group have high chances of having coronary heart disease in 10 years

```
In [150]: #Read the csv
df = pd.read_csv('framingham.csv')

#Filter and make a new dataframe for those who has chances of having coronary heart disease
coronary_heart_disease = df.loc[df['TenYearCHD'] == 1]
coronary_heart_disease
```

```
Out[150]:
```

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | c |
|------|------|-----|-----------|---------------|------------|--------|-----------------|--------------|-----|
| 3 | 0 | 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | 1 | |
| 6 | 0 | 63 | 1.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 15 | 0 | 38 | 2.0 | 1 | 20.0 | 0.0 | 0 | 1 | |
| 17 | 0 | 46 | 2.0 | 1 | 20.0 | 0.0 | 0 | 0 | |
| 25 | 1 | 47 | 4.0 | 1 | 20.0 | 0.0 | 0 | 0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4221 | 1 | 50 | 1.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 4223 | 1 | 56 | 4.0 | 0 | 0.0 | 1.0 | 0 | 1 | |
| 4226 | 1 | 58 | 1.0 | 0 | 0.0 | 0.0 | 0 | 0 | |
| 4232 | 1 | 68 | 1.0 | 0 | 0.0 | 0.0 | 0 | 1 | |
| 4233 | 1 | 50 | 1.0 | 1 | 1.0 | 0.0 | 0 | 1 | |

644 rows × 16 columns



```
In [151]: #Group by age column and count the rows of TenYearCHD column  
group_age = coronary_heart_disease.groupby('age')['TenYearCHD'].count().reset_index  
group_age
```

Out[151]:

| | age | TenYearCHD |
|----|-----|------------|
| 0 | 35 | 2 |
| 1 | 36 | 3 |
| 2 | 37 | 4 |
| 3 | 38 | 8 |
| 4 | 39 | 6 |
| 5 | 40 | 15 |
| 6 | 41 | 11 |
| 7 | 42 | 14 |
| 8 | 43 | 13 |
| 9 | 44 | 16 |
| 10 | 45 | 14 |
| 11 | 46 | 16 |
| 12 | 47 | 23 |
| 13 | 48 | 21 |
| 14 | 49 | 24 |
| 15 | 50 | 23 |
| 16 | 51 | 29 |
| 17 | 52 | 32 |
| 18 | 53 | 23 |
| 19 | 54 | 18 |
| 20 | 55 | 24 |
| 21 | 56 | 27 |
| 22 | 57 | 26 |
| 23 | 58 | 31 |
| 24 | 59 | 30 |
| 25 | 60 | 26 |
| 26 | 61 | 25 |
| 27 | 62 | 25 |
| 28 | 63 | 32 |
| 29 | 64 | 21 |
| 30 | 65 | 20 |
| 31 | 66 | 15 |
| 32 | 67 | 17 |

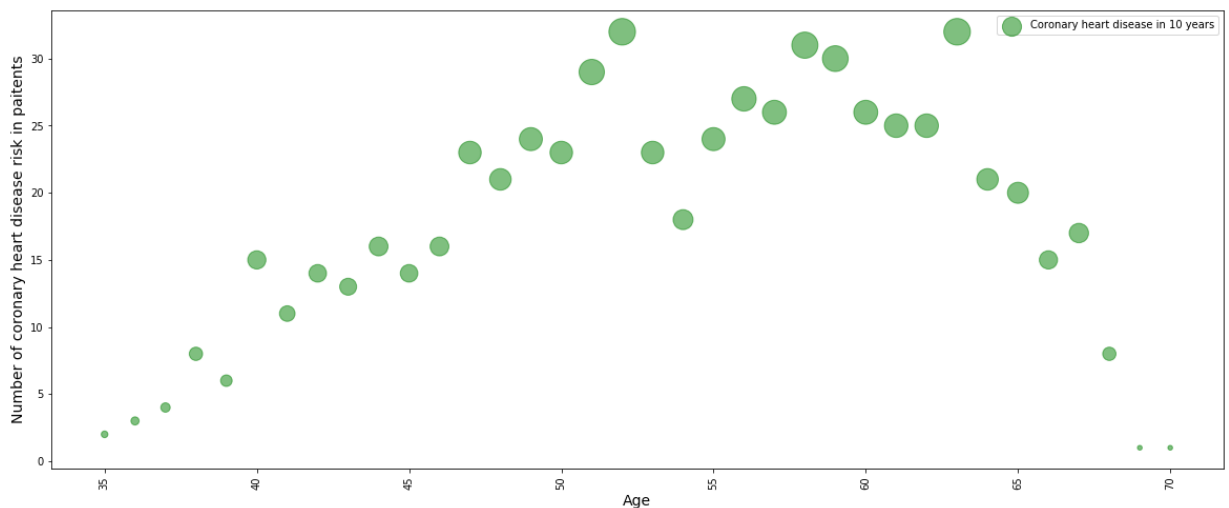
| | age | TenYearCHD |
|----|-----|------------|
| 33 | 68 | 8 |
| 34 | 69 | 1 |
| 35 | 70 | 1 |

In [154]: *#Plot a bubble graph to show total number of people having a chance of coronary heart disease in 10 years*
plt.figure(figsize=(20, 8))

```
plt.scatter(group_age['age'], group_age['TenYearCHD'],
            color='green', label='Coronary heart disease in 10 years',
            alpha=0.5,
            s = group_age['TenYearCHD']*20)
```

```
plt.xticks(rotation='vertical')
plt.legend()
plt.xlabel("Age", size=14)
plt.ylabel("Number of coronary heart disease risk in patients", size=14)
```

Out[154]: Text(0, 0.5, 'Number of coronary heart disease risk in patients')



Conslusion - The age group between 50 to 65 has a higher chance of having coronary heart disease in 10 years.

In []: