In [1]:

```python
print("Name : Jas")
print("We will be cleaning the big data and make a comparison to show who has a healthier h
print("Also we will be deriviring which age group has the high chances of coronary heart di
```

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

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

In [2]:

```python
#Import libraries

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
#read the csv

dataframe=pd.read_csv('framingham.csv')
dataframe
```

Out[2]:

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

4238 rows × 16 columns

In [3]:

```python
#Filter and make a new dataframe for non smokers

non_smokers_dataframe=dataframe.loc[dataframe['currentSmoker']==0]
non_smokers_dataframe
```

Out[3]:

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

2144 rows × 16 columns

In [4]:

```
#Group by age column and find average heart rate at different age

dataframe_average_heart_beat_for_non_smokers=non_smokers_dataframe.groupby('age')['heartRat
dataframe_average_heart_beat_for_non_smokers
```

Out[4]:

	age	heartRate
0	33	76.000000
1	34	74.000000
2	35	74.789474
3	36	74.216216
4	37	76.027778
5	38	72.232143
6	39	76.239437
7	40	77.885714
8	41	73.083333
9	42	75.042857
10	43	75.050000
11	44	73.746032
12	45	77.333333
13	46	77.038961
14	47	75.173077
15	48	75.176471
16	49	73.868852
17	50	75.791045
18	51	74.200000
19	52	76.560440
20	53	77.125000
21	54	74.437500
22	55	74.305263
23	56	73.397059
24	57	74.027778
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

	age	heartRate
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 [5]:

```python
#Filter and make a new dataframe for smokers

smokers_dataframe=dataframe.loc[dataframe['currentSmoker']==1]
smokers_dataframe
```

Out[5]:

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

2094 rows × 16 columns

In [6]:

```
#Group by age column and find average heart rate at different age

dataframe_average_heart_beat_for_smokers=smokers_dataframe.groupby('age')['heartRate'].mean
dataframe_average_heart_beat_for_smokers
```

Out[6]:

	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 [7]:

```python
#Plot a line graph to show the heart rate of smokers vs non smokers

plt.figure(figsize=(25,25))

label=dataframe_average_heart_beat_for_non_smokers['age']
value=dataframe_average_heart_beat_for_non_smokers['heartRate']

plt.plot(label,value,label="Non Smokers",linewidth=3.0)

label=dataframe_average_heart_beat_for_smokers['age']
value=dataframe_average_heart_beat_for_smokers['heartRate']

plt.plot(label,value,label="Smokers",linewidth=3.0)

plt.xlabel('Age',size=30)
plt.ylabel('Heart Rate',size=30)

plt.title("Difference Between Heart Rate of Smokers and Non Smokers",fontsize=50)
plt.legend()
plt.show()
```
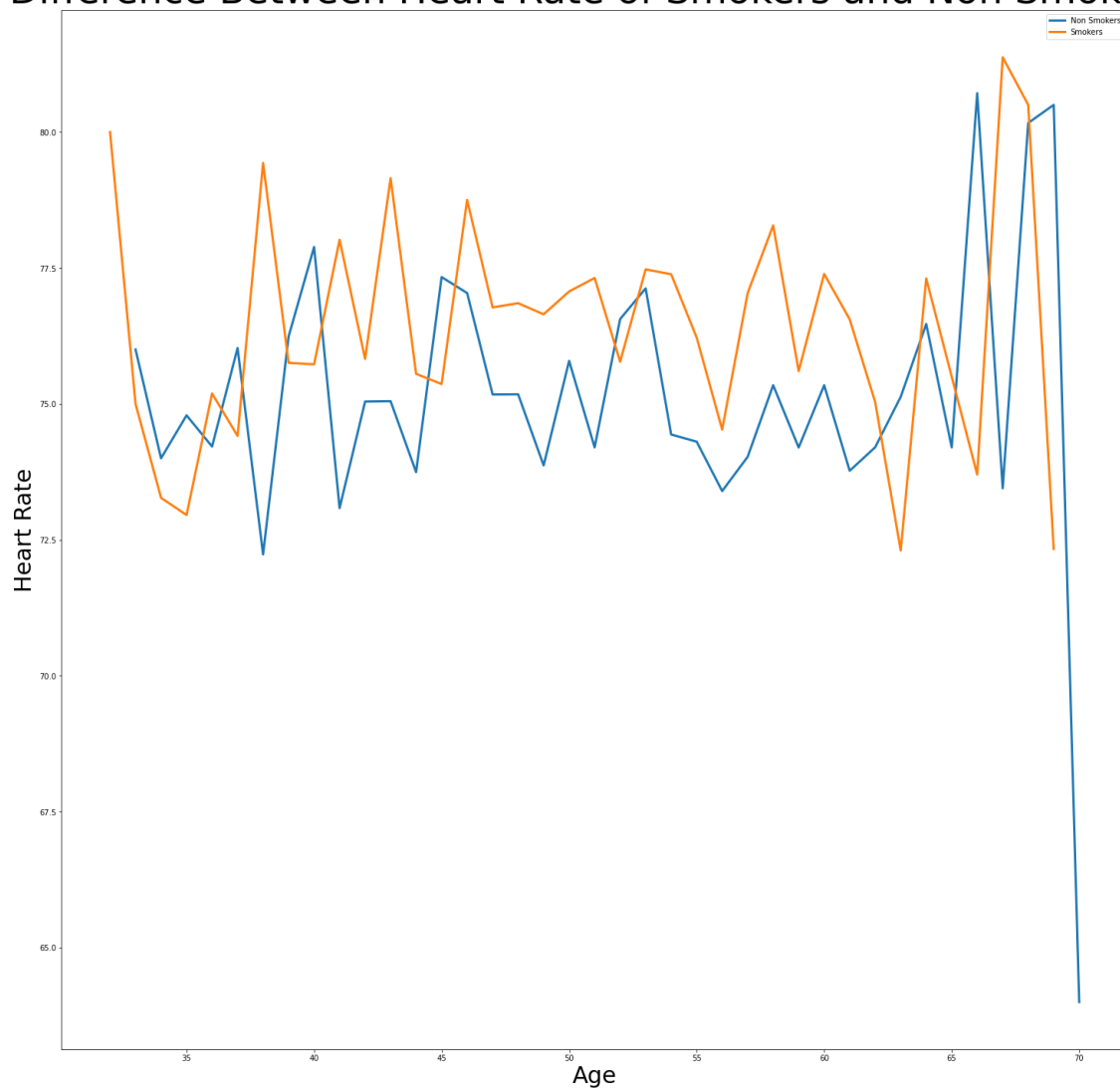

Conslusion - The heart beat of smokers is slightly more than that of non smokers.

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

In [8]:

```
#Read the csv

dataframe_coronary_disease=pd.read_csv('framingham.csv')
dataframe_coronary_disease
#Filter and make a new dataframe for those who has chances of having coronary heart disease

coronary_heart_disease_dataframe=dataframe_coronary_disease.loc[dataframe_coronary_disease[
coronary_heart_disease_dataframe
```

Out[8]:

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

644 rows × 16 columns

In [9]:
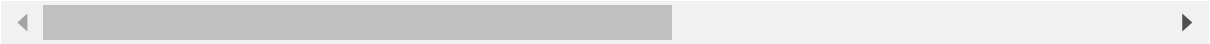
```python
#Group by age column and count the rows of TenYearCHD column

coronary_heart_disease_dataframe_10=coronary_heart_disease_dataframe.groupby('age')['TenYea
coronary_heart_disease_dataframe_10
```

Out[9]:

	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

	age	TenYearCHD
32	67	17
33	68	8
34	69	1
35	70	1
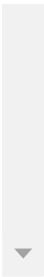
In [25]:

```python
#Plot a line graph to show total number of people having a chance of coronary heart disease

plt.figure(figsize=(25,25))

rgb=np.random.rand(3)
rgb

alpha_value=np.random.rand()
alpha_value

plt.xlabel("Age",size=30)
plt.xticks(rotation='vertical')
plt.ylabel("",size=30)

plt.scatter(coronary_heart_disease_dataframe_10['age'],coronary_heart_disease_dataframe_10[
plt.legend()
plt.title("High chances of coronary heart disease",size=50)
plt.show()
```
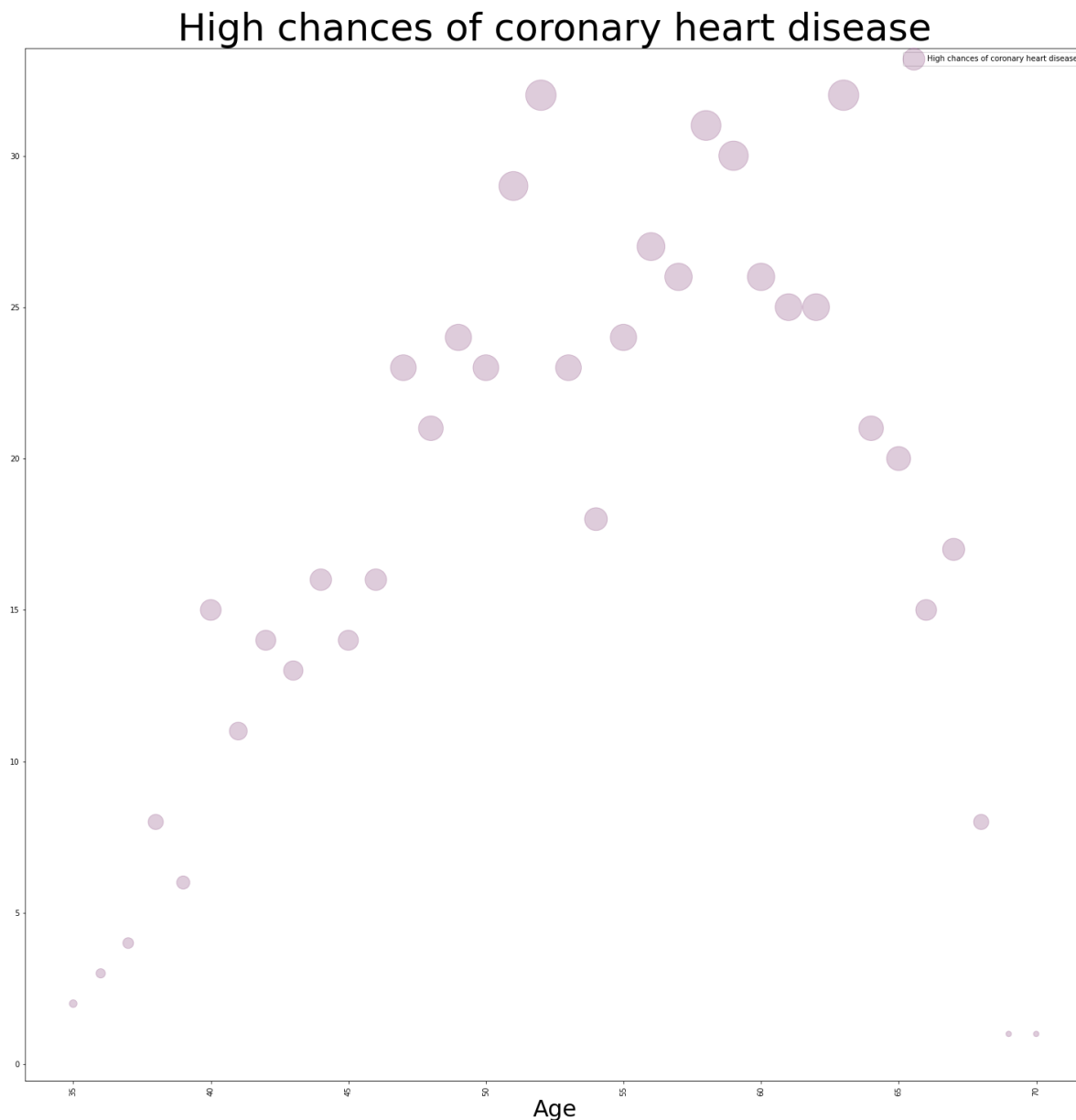
High chances of coronary heart disease

Conslusion - The highest chance of a coronary disease is at the age of 58 and 63

In [ ]: