Minimum Viable Product (MVP)

Cardiovascular Disease dataset

The goal of this project is to analyze Cardiovascular Disease dataset to find which variables are related to the disease. Then we will use different machine learning models to predict whether the patient has cardiovascular disease or not.
The dataset contains information about patients doing cardiovascular disease examination.

Let's start with Loading and Displaying data:

In [2]:
```
#laoding data
df = pd.read_csv('./downloads/cardio_train.txt.csv' , sep=';')
```

In [3]:
```
# Let us now begin first with finding some quick descriptive stats about our data.
df.head(3)
```

Out[3]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1

In [5]:
```
df.shape
```

Out[5]:
```
(70000, 13)
```

In [8]:
```
df.columns
```

Out[8]:
```
Index(['id', 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo',
       'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'cardio'],
      dtype='object')
```

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

Out[6]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.

Data Understanding (EDA & Visualizations)

In [38]:
```
# Identifying missing values.
df.isnull().head(3)
```

Out[38]:

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False

In [36]:
```
df.isnull().sum()
```

Out[36]:
```
age            0
gender         0
height         0
weight         0
ap_hi          0
ap_lo          0
cholesterol    0
gluc           0
smoke          0
alco           0
active         0
cardio         0
dtype: int64
```

In [8]:
```
del df['id'] #Lets drop id since it does not provide any additional information
```

In [9]:
```
df.head(2)
```

Out[9]:

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1

From above we see that we don't have any missing values

How many Smoker and Non-Smoker in the dataset?

:
```
num = df["smoke"].value_counts("0")
num
```

:
```
0    0.911317
1    0.088683
Name: smoke, dtype: float64
```

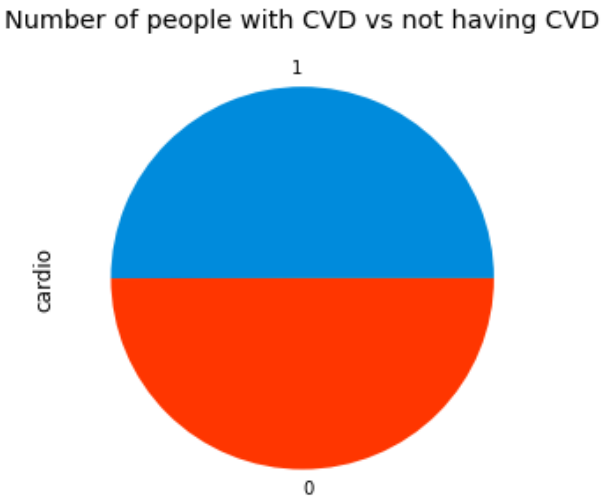
:
```
non_smoker = 70000 * 0.911871
smoker = 70000 * 0.088129

print("There are " + str(non_smoker) + " Non-Smokers and " + str(smoker) + " Smokers in the dataset.")
```

```
There are 63830.97 Non-Smokers and 6169.03 Smokers in the dataset.
```

To answer the question how many people in the dataset have a cardiovascular disease? we use a pie chart below

```
df['cardio'].value_counts().plot.pie(figsize=(5, 5))
plt.title("Number of people with CVD vs not having CVD")
plt.show()
```



Concluding the chart, we've found that:

The percentage of people with cardiovascular diseases is 50%.

We used Scatter matrix to visualize our data