In [2]:
```
import pandas as pd
```

In [3]:
```
import numpy as np
```

In [4]:
```
df = sns.load_dataset("archive")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3115619703.py in <module>
----> 1 df = sns.load_dataset("archive")

NameError: name 'sns' is not defined
```

In [5]:
```
import seaborn as sns
```

In [6]:
```
df = sns.load_dataset("archive")
```

```
---------------------------------------------------------------------------
TimeoutError                              Traceback (most recent call last)
~\anaconda3\lib\urllib\request.py in do_open(self, http_class, req, **http_conn_args)
   1345             try:
-> 1346                 h.request(req.get_method(), req.selector, req.data, headers,
   1347                           encode_chunked=req.has_header('Transfer-encoding'))

~\anaconda3\lib\http\client.py in request(self, method, url, body, headers, encode_ch
unked)
   1278         """Send a complete request to the server."""
-> 1279         self._send_request(method, url, body, headers, encode_chunked)
   1280

~\anaconda3\lib\http\client.py in _send_request(self, method, url, body, headers, enc
ode_chunked)
   1324             body = _encode(body, 'body')
-> 1325         self.endheaders(body, encode_chunked=encode_chunked)
   1326

~\anaconda3\lib\http\client.py in endheaders(self, message_body, encode_chunked)
   1273             raise CannotSendHeader()
-> 1274         self._send_output(message_body, encode_chunked=encode_chunked)
   1275

~\anaconda3\lib\http\client.py in _send_output(self, message_body, encode_chunked)
   1033         del self._buffer[:]
-> 1034         self.send(msg)
   1035

~\anaconda3\lib\http\client.py in send(self, data)
    973             if self.auto_open:
--> 974                 self.connect()
    975             else:

~\anaconda3\lib\http\client.py in connect(self)
   1440
-> 1441             super().connect()
   1442

~\anaconda3\lib\http\client.py in connect(self)
```

```
    944         """Connect to the host and port specified in __init__."""
--> 945         self.sock = self._create_connection(
    946             (self.host,self.port), self.timeout, self.source_address)

~\anaconda3\lib\socket.py in create_connection(address, timeout, source_address)
    843         try:
--> 844             raise err
    845         finally:

~\anaconda3\lib\socket.py in create_connection(address, timeout, source_address)
    831                 sock.bind(source_address)
--> 832             sock.connect(sa)
    833             # Break explicitly a reference cycle

TimeoutError: [WinError 10060] A connection attempt failed because the connected part
y did not properly respond after a period of time, or established connection failed b
ecause connected host has failed to respond

During handling of the above exception, another exception occurred:

URLError                                  Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3115619703.py in <module>
----> 1 df = sns.load_dataset("archive")

~\anaconda3\lib\site-packages\seaborn\utils.py in load_dataset(name, cache, data_hom
e, **kws)
    592         cache_path = os.path.join(get_data_home(data_home), os.path.basename
(url))
    593         if not os.path.exists(cache_path):
--> 594             if name not in get_dataset_names():
    595                 raise ValueError(f"'{name}' is not one of the example dataset
s.")
    596             urlretrieve(url, cache_path)

~\anaconda3\lib\site-packages\seaborn\utils.py in get_dataset_names()
    518     """
    519     url = "https://github.com/mwaskom/seaborn-data"
--> 520     with urlopen(url) as resp:
    521         html = resp.read()
    522

~\anaconda3\lib\urllib\request.py in urlopen(url, data, timeout, cafile, capath, cade
fault, context)
    212     else:
    213         opener = _opener
--> 214     return opener.open(url, data, timeout)
    215
    216 def install_opener(opener):

~\anaconda3\lib\urllib\request.py in open(self, fullurl, data, timeout)
    515
    516         sys.audit('urllib.Request', req.full_url, req.data, req.headers, req.
get_method())
--> 517         response = self._open(req, data)
    518
    519         # post-process response

~\anaconda3\lib\urllib\request.py in _open(self, req, data)
    532
    533         protocol = req.type
--> 534         result = self._call_chain(self.handle_open, protocol, protocol +
    535                                   '_open', req)
    536         if result:
```

```
~\anaconda3\lib\urllib\request.py in _call_chain(self, chain, kind, meth_name, *args)
    492         for handler in handlers:
    493             func = getattr(handler, meth_name)
--> 494             result = func(*args)
    495             if result is not None:
    496                 return result

~\anaconda3\lib\urllib\request.py in https_open(self, req)
   1387
   1388         def https_open(self, req):
-> 1389             return self.do_open(http.client.HTTPSConnection, req,
   1390                 context=self._context, check_hostname=self._check_hostname)
   1391

~\anaconda3\lib\urllib\request.py in do_open(self, http_class, req, **http_conn_args)
   1347                             encode_chunked=req.has_header('Transfer-encoding'))
   1348             except OSError as err: # timeout error
-> 1349                 raise URLError(err)
   1350             r = h.getresponse()
   1351         except:

URLError: <urlopen error [WinError 10060] A connection attempt failed because the con
nected party did not properly respond after a period of time, or established connecti
on failed because connected host has failed to respond>
```

In [7]:
```
df = sns.load_dataset("diabetes_data")
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3038040564.py in <module>
----> 1 df = sns.load_dataset("diabetes_data")

~\anaconda3\lib\site-packages\seaborn\utils.py in load_dataset(name, cache, data_hom
e, **kws)
    593         if not os.path.exists(cache_path):
    594             if name not in get_dataset_names():
--> 595                 raise ValueError(f"'{name}' is not one of the example dataset
s.")
    596             urlretrieve(url, cache_path)
    597         full_path = cache_path

ValueError: 'diabetes_data' is not one of the example datasets.
```

In [10]:
```
file_path = "diabetes_data.csv"
```

In [11]:
```
df = pd.read_csv(r"C:\Users\user\Desktop\DAFINAL PROJECT\archive\diabetes_data.csv")
```

In [12]:
```
df
```

Out[12]:

	Age	Sex	HighChol	CholCheck	BMI	Smoker	HeartDiseaseorAttack	PhysActivity	Fruits	V
0	4.0	1.0	0.0	1.0	26.0	0.0	0.0	1.0	0.0	
1	12.0	1.0	1.0	1.0	26.0	1.0	0.0	0.0	1.0	
2	13.0	1.0	0.0	1.0	26.0	0.0	0.0	1.0	1.0	
3	11.0	1.0	1.0	1.0	28.0	1.0	0.0	1.0	1.0	
4	8.0	0.0	0.0	1.0	29.0	1.0	0.0	1.0	1.0	

	Age	Sex	HighChol	CholCheck	BMI	Smoker	HeartDiseaseorAttack	PhysActivity	Fruits	V
...	...	...	...	...	...	...	...	...	...	
70687	6.0	0.0	1.0	1.0	37.0	0.0	0.0	0.0	0.0	
70688	10.0	1.0	1.0	1.0	29.0	1.0	1.0	0.0	1.0	
70689	13.0	0.0	1.0	1.0	25.0	0.0	1.0	0.0	1.0	
70690	11.0	0.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	
70691	9.0	0.0	1.0	1.0	25.0	0.0	1.0	1.0	1.0	

70692 rows × 18 columns

In [13]:
```
file_path = "hypertension_data.csv"
```

In [14]:
```
df = pd.read_csv(r"C:\Users\user\Desktop\DAFINAL PROJECT\archive\hypertension_data.c
```

In [15]:
```
df
```

Out[15]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	57.0	1.0	3	145	233	1	0	150	0	2.3	0	0	1	1
1	64.0	0.0	2	130	250	0	1	187	0	3.5	0	0	2	1
2	52.0	1.0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56.0	0.0	1	120	236	0	1	178	0	0.8	2	0	2	1
4	66.0	0.0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26078	72.0	0.0	0	138	294	1	1	106	0	1.9	1	3	2	0
26079	60.0	1.0	0	144	200	0	0	126	1	0.9	1	0	3	0
26080	68.0	1.0	0	100	234	0	1	156	0	0.1	2	1	3	0
26081	67.0	1.0	1	154	232	0	0	164	0	0.0	2	1	2	0
26082	67.0	0.0	1	130	236	0	0	174	0	0.0	1	1	2	0

26083 rows × 14 columns

In [16]:
```
file_path = "stroke_data.csv"
```

In [17]:
```
df = pd.read_csv(r"C:\Users\user\Desktop\DAFINAL PROJECT\archive\stroke_data.csv")
```

In [18]:
```
df
```

Out[18]:

	sex	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose
0	1.0	63.0	0	1	1	4	1	2
1	1.0	42.0	0	1	1	4	0	1
2	0.0	61.0	0	0	1	4	1	1
3	1.0	41.0	1	0	1	3	0	1
4	1.0	85.0	0	0	1	4	1	1
...	...	...	...	...	...	...	...	
40905	1.0	38.0	0	0	0	4	1	1
40906	0.0	53.0	0	0	1	4	0	
40907	1.0	32.0	0	0	1	2	0	2
40908	1.0	42.0	0	0	1	3	0	2
40909	1.0	35.0	0	0	0	4	0	

40910 rows × 11 columns

In [19]:
```
print(df.head())
```

```
   sex   age  hypertension  heart_disease  ever_married  work_type  \
0  1.0  63.0             0              1             1          4
1  1.0  42.0             0              1             1          4
2  0.0  61.0             0              0             1          4
3  1.0  41.0             1              0             1          3
4  1.0  85.0             0              0             1          4

   Residence_type  avg_glucose_level   bmi  smoking_status  stroke
0               1             228.69  36.6               1       1
1               0             105.92  32.5               0       1
2               1             171.23  34.4               1       1
3               0             174.12  24.0               0       1
4               1             186.21  29.0               1       1
```

In [20]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40910 entries, 0 to 40909
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sex                40907 non-null  float64
 1   age                40910 non-null  float64
 2   hypertension       40910 non-null  int64
 3   heart_disease      40910 non-null  int64
 4   ever_married       40910 non-null  int64
 5   work_type          40910 non-null  int64
 6   Residence_type     40910 non-null  int64
 7   avg_glucose_level  40910 non-null  float64
 8   bmi                40910 non-null  float64
 9   smoking_status     40910 non-null  int64
 10  stroke             40910 non-null  int64
dtypes: float64(4), int64(7)
memory usage: 3.4 MB
None
```

In [21]:
```
print(df.isnull().sum())
```

```
sex                 3
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
dtype: int64
```

In [22]:
```
df = df.dropna()
```

In [23]:
```
print(df.duplicated().sum())
```

```
0
```

In [24]:
```
df = df.drop_duplicates()
```

In [25]:
```
df['date_column'] = pd.to_datetime(df['date_column'])
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
~\anaconda3\lib\site-packages\pandas\core\indexes\base.py in get_loc(self, key, metho
d, tolerance)
   3360             try:
-> 3361                 return self._engine.get_loc(casted_key)
   3362             except KeyError as err:

~\anaconda3\lib\site-packages\pandas\_libs\index.pyx in pandas._libs.index.IndexEngin
e.get_loc()

~\anaconda3\lib\site-packages\pandas\_libs\index.pyx in pandas._libs.index.IndexEngin
e.get_loc()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.g
et_item()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.g
et_item()

KeyError: 'date_column'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3411264587.py in <module>
----> 1 df['date_column'] = pd.to_datetime(df['date_column'])

~\anaconda3\lib\site-packages\pandas\core\frame.py in __getitem__(self, key)
   3456             if self.columns.nlevels > 1:
   3457                 return self._getitem_multilevel(key)
-> 3458             indexer = self.columns.get_loc(key)
   3459             if is_integer(indexer):
```

```
   3460                 indexer = [indexer]

~\anaconda3\lib\site-packages\pandas\core\indexes\base.py in get_loc(self, key, metho
d, tolerance)
   3361                 return self._engine.get_loc(casted_key)
   3362             except KeyError as err:
-> 3363                 raise KeyError(key) from err
   3364
   3365         if is_scalar(key) and isna(key) and not self.hasnans:

KeyError: 'date_column'
```

In [26]:

```
print(df.describe()
```

```
  File "C:\Users\user\AppData\Local\Temp/ipykernel_2560/4270986709.py", line 1
    print(df.describe()
                       ^
SyntaxError: unexpected EOF while parsing
```

In [27]:

```
print(df.describe())
```

```
                sex           age  hypertension  heart_disease  ever_married  \
count  40907.000000  40907.000000  40907.000000   40907.000000  40907.000000
mean       0.555162     51.327303      0.213851       0.127729      0.821326
std        0.496954     21.624171      0.410028       0.333792      0.383083
min        0.000000     -9.000000      0.000000       0.000000      0.000000
25%        0.000000     35.000000      0.000000       0.000000      1.000000
50%        1.000000     52.000000      0.000000       0.000000      1.000000
75%        1.000000     68.000000      0.000000       0.000000      1.000000
max        1.000000    103.000000      1.000000       1.000000      1.000000

          work_type  Residence_type  avg_glucose_level           bmi  \
count  40907.000000    40907.000000       40907.000000  40907.000000
mean       3.461095        0.514851         122.079679     30.406488
std        0.780934        0.499786          57.561951      6.835305
min        0.000000        0.000000          55.120000     11.500000
25%        3.000000        0.000000          78.750000     25.900000
50%        4.000000        1.000000          97.920000     29.400000
75%        4.000000        1.000000         167.590000     34.100000
max        4.000000        1.000000         271.740000     92.000000

       smoking_status        stroke
count    40907.000000  40907.000000
mean         0.488572      0.500159
std          0.499875      0.500006
min          0.000000      0.000000
25%          0.000000      0.000000
50%          0.000000      1.000000
75%          1.000000      1.000000
max          1.000000      1.000000
```

In [28]:

```
import matplotlib.pylot as plt
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3792379548.py in <module>
----> 1 import matplotlib.pylot as plt

ModuleNotFoundError: No module named 'matplotlib.pylot'
```

In [29]:

```
pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-package
s (3.4.3)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\user\anaconda3\lib\si
te-packages (from matplotlib) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\user\anaconda3\lib\site-p
ackages (from matplotlib) (3.0.4)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\user\anaconda3\lib\site-
packages (from matplotlib) (1.3.1)
Requirement already satisfied: numpy>=1.16 in c:\users\user\anaconda3\lib\site-packag
es (from matplotlib) (1.20.3)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packa
ges (from matplotlib) (0.10.0)
Requirement already satisfied: pillow>=6.2.0 in c:\users\user\anaconda3\lib\site-pack
ages (from matplotlib) (8.4.0)
Requirement already satisfied: six in c:\users\user\anaconda3\lib\site-packages (from
cycler>=0.10->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [30]:
```
import matplotlib.pylot as plt
```

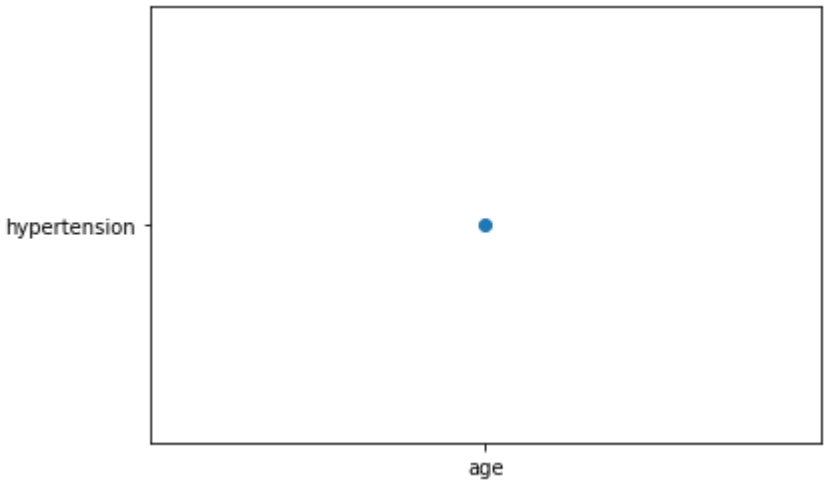
```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_2560/3792379548.py in <module>
----> 1 import matplotlib.pylot as plt

ModuleNotFoundError: No module named 'matplotlib.pylot'
```

In [31]:
```
import matplotlib.pyplot as plt
```

In [32]:
```
plt.scatter(['age'],['hypertension'])
```

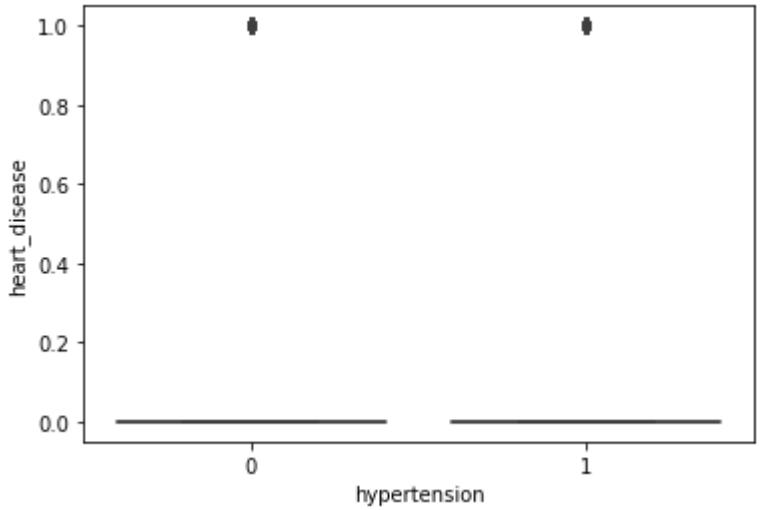
Out[32]:
```
<matplotlib.collections.PathCollection at 0x1af8187e370>
```



In [33]:
```
df1 = df.dropna(subset=['hypertension', 'heart_disease'])
```

In [34]:
```
sns.boxplot(data=df1,x='hypertension',y='heart_disease')
```

Out[34]:
```
<AxesSubplot:xlabel='hypertension', ylabel='heart_disease'>
```



In [35]:
```
df_corr = df[['age', 'stroke']]
```

In [36]:
```
corr_matrix = df_corr.corr()
```

In [37]:
```
sns.heatmap(corr_matrix, annot=True, cmap='YlGnBu', linewidths=0.5, fmt=".2f")
```

Out[37]: <AxesSubplot:>

In [ ]: