In [1]:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df=pd.read_csv(r"chd.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totCh
0	1	39	0	0.0	0.0	0	0	0	195
1	0	46	0	0.0	0.0	0	0	0	250
2	1	48	1	20.0	0.0	0	0	0	245
3	0	61	1	30.0	0.0	0	1	0	225
4	0	46	1	23.0	0.0	0	0	0	285

In [4]:

```
df.dtypes
```

Out[4]:

```
male                  int64
age                   int64
currentSmoker         int64
cigsPerDay          float64
BPMeds              float64
prevalentStroke       int64
prevalentHyp          int64
diabetes              int64
totChol             float64
sysBP                 int64
diaBP                 int64
BMI                 float64
heartRate           float64
glucose             float64
TenYearCHD            int64
dtype: object
```

In [5]:

```
df.info
```

Out[5]:

```
<bound method DataFrame.info of       male  age  currentSmoker  cigsPerDay  BPM
eds  prevalentStroke  \
0        1   39              0         0.0     0.0               0
1        0   46              0         0.0     0.0               0
2        1   48              1        20.0     0.0               0
3        0   61              1        30.0     0.0               0
4        0   46              1        23.0     0.0               0
...    ...  ...            ...         ...     ...             ...
4233     1   50              1         1.0     0.0               0
4234     1   51              1        43.0     0.0               0
4235     0   48              1        20.0     NaN               0
4236     0   44              1        15.0     0.0               0
4237     0   52              0         0.0     0.0               0

      prevalentHyp  diabetes  totChol  sysBP  diaBP   BMI  heartRate  glucose
\
0                0         0    195.0    106     70  27.0       80.0     77.0
1                0         0    250.0    121     81  29.0       95.0     76.0
2                0         0    245.0    128     80  25.0       75.0     70.0
3                1         0    225.0    150     95  29.0       65.0    103.0
4                0         0    285.0    130     84  23.0       85.0     85.0
...            ...       ...      ...    ...    ...   ...        ...      ...
4233             1         0    313.0    179     92  26.0       66.0     86.0
4234             0         0    207.0    127     80  20.0       65.0     68.0
4235             0         0    248.0    131     72  22.0       84.0     86.0
4236             0         0    210.0    127     87  19.0       86.0      NaN
4237             0         0    269.0    134     83  21.0       80.0    107.0

      TenYearCHD
0              0
1              0
2              0
3              1
4              0
...          ...
4233           1
4234           0
4235           0
4236           0
4237           0

[4238 rows x 15 columns]>
```

In [6]:

```
df.memory_usage()
```

Out[6]:

```
Index                128
male               33904
age                33904
currentSmoker      33904
cigsPerDay         33904
BPMeds             33904
prevalentStroke    33904
prevalentHyp       33904
diabetes           33904
totChol            33904
sysBP              33904
diaBP              33904
BMI                33904
heartRate          33904
glucose            33904
TenYearCHD         33904
dtype: int64
```

In [7]:

```
df.memory_usage().sum()
```

Out[7]:

```
508688
```

In [8]:

```
df.describe()
```

Out[8]:

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	preval
count	4238.000000	4238.000000	4238.000000	4209.000000	4185.000000	4238.000000	4238.
mean	0.429212	49.584946	0.494101	9.003089	0.029630	0.005899	0.
std	0.495022	8.572160	0.500024	11.920094	0.169584	0.076587	0.
min	0.000000	32.000000	0.000000	0.000000	0.000000	0.000000	0.
25%	0.000000	42.000000	0.000000	0.000000	0.000000	0.000000	0.
50%	0.000000	49.000000	0.000000	0.000000	0.000000	0.000000	0.
75%	1.000000	56.000000	1.000000	20.000000	0.000000	0.000000	1.
max	1.000000	70.000000	1.000000	70.000000	1.000000	1.000000	1.

In [9]:
```
df.mean()
```

Out[9]:
```
male                  0.429212
age                  49.584946
currentSmoker         0.494101
cigsPerDay            9.003089
BPMeds                0.029630
prevalentStroke       0.005899
prevalentHyp          0.310524
diabetes              0.025720
totChol             236.721585
sysBP               132.449976
diaBP                82.974280
BMI                  25.808722
heartRate            75.878924
glucose              81.966753
TenYearCHD            0.151958
dtype: float64
```

In [10]:
```
df['BMI'].mean()
```

Out[10]:
```
25.80872244607727
```

In [11]:
```
df.var()
```

Out[11]:
```
male                   0.245047
age                   73.481926
currentSmoker          0.250024
cigsPerDay           142.088631
BPMeds                 0.028759
prevalentStroke        0.005866
prevalentHyp           0.214149
diabetes               0.025064
totChol             1988.297915
sysBP                485.617393
diaBP                141.778191
BMI                   16.743158
heartRate            144.639020
glucose              574.081513
TenYearCHD             0.128898
dtype: float64
```

```
In [12]: df.skew()
```

```
Out[12]: male                0.286135
         age                 0.228146
         currentSmoker       0.023606
         cigsPerDay          1.247910
         BPMeds              5.550010
         prevalentStroke    12.909062
         prevalentHyp        0.819278
         diabetes            5.994378
         totChol             0.871422
         sysBP               1.143799
         diaBP               0.714524
         BMI                 0.984374
         heartRate           0.644482
         glucose             6.213402
         TenYearCHD          1.939741
         dtype: float64
```

```
In [13]: df.kurtosis()
```

```
Out[13]: male                -1.919033
         age                 -0.989636
         currentSmoker       -2.000387
         cigsPerDay           1.023356
         BPMeds              28.816384
         prevalentStroke    164.721624
         prevalentHyp        -1.329411
         diabetes            33.948587
         totChol              4.131582
         sysBP                2.146845
         diaBP                1.280286
         BMI                  2.658429
         heartRate            0.907483
         glucose             58.674278
         TenYearCHD           1.763428
         dtype: float64
```

In [14]:

```
df.min()
```

Out[14]:

```
male                 0.0
age                 32.0
currentSmoker        0.0
cigsPerDay           0.0
BPMeds               0.0
prevalentStroke      0.0
prevalentHyp         0.0
diabetes             0.0
totChol            107.0
sysBP               84.0
diaBP               48.0
BMI                 16.0
heartRate           44.0
glucose             40.0
TenYearCHD           0.0
dtype: float64
```

In [15]:

```
df.max()
```

Out[15]:

```
male                 1.0
age                 70.0
currentSmoker        1.0
cigsPerDay          70.0
BPMeds               1.0
prevalentStroke      1.0
prevalentHyp         1.0
diabetes             1.0
totChol            696.0
sysBP              295.0
diaBP              143.0
BMI                 57.0
heartRate          143.0
glucose            394.0
TenYearCHD           1.0
dtype: float64
```

In [16]:
```
df.median()
```

Out[16]:
```
male                 0.0
age                 49.0
currentSmoker        0.0
cigsPerDay           0.0
BPMeds               0.0
prevalentStroke      0.0
prevalentHyp         0.0
diabetes             0.0
totChol            234.0
sysBP              128.0
diaBP               82.0
BMI                 25.0
heartRate           75.0
glucose             78.0
TenYearCHD           0.0
dtype: float64
```

In [17]:
```
df.corr()
```

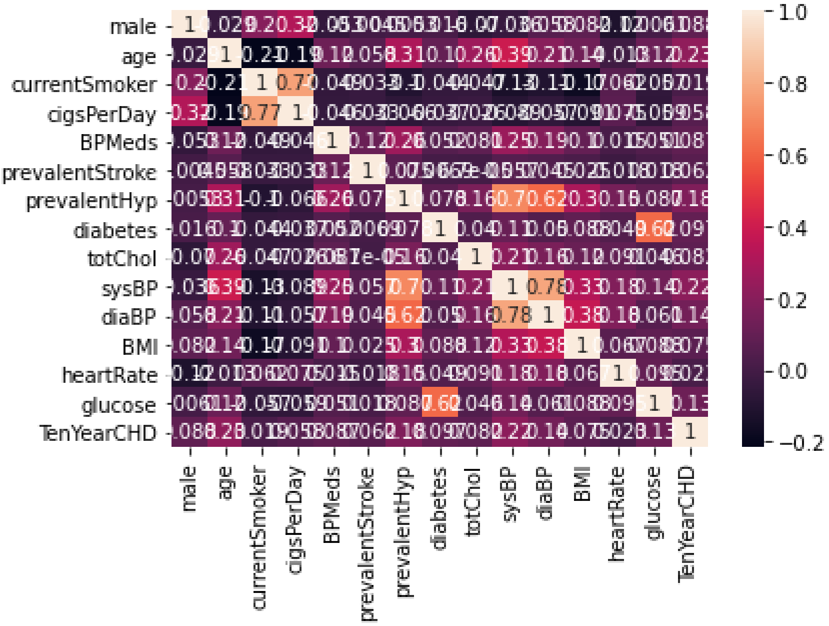
Out[17]:

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	preva
male	1.000000	-0.028979	0.197596	0.317930	-0.052506	-0.004546	(
age	-0.028979	1.000000	-0.213748	-0.192791	0.122995	0.057655	(
currentSmoker	0.197596	-0.213748	1.000000	0.769690	-0.048938	-0.032988	-(
cigsPerDay	0.317930	-0.192791	0.769690	1.000000	-0.046134	-0.032707	-(
BPMeds	-0.052506	0.122995	-0.048938	-0.046134	1.000000	0.117365	(
prevalentStroke	-0.004546	0.057655	-0.032988	-0.032707	0.117365	1.000000	(
prevalentHyp	0.005313	0.307194	-0.103260	-0.066146	0.261187	0.074830	1
diabetes	0.015708	0.101258	-0.044295	-0.037067	0.052047	0.006949	(
totChol	-0.070322	0.262131	-0.046562	-0.026320	0.080558	0.000067	(
sysBP	-0.035969	0.394061	-0.130298	-0.088785	0.253834	0.056741	(
diaBP	0.057892	0.205481	-0.108067	-0.056936	0.193806	0.044941	(
BMI	0.082145	0.135356	-0.166717	-0.090740	0.100340	0.024704	(
heartRate	-0.116620	-0.012823	0.062356	0.075157	0.015233	-0.017676	(
glucose	0.006083	0.122256	-0.056826	-0.058960	0.051176	0.018431	(
TenYearCHD	0.088428	0.225256	0.019456	0.057884	0.087489	0.061810	(

In [18]:
```
import seaborn as sns
```

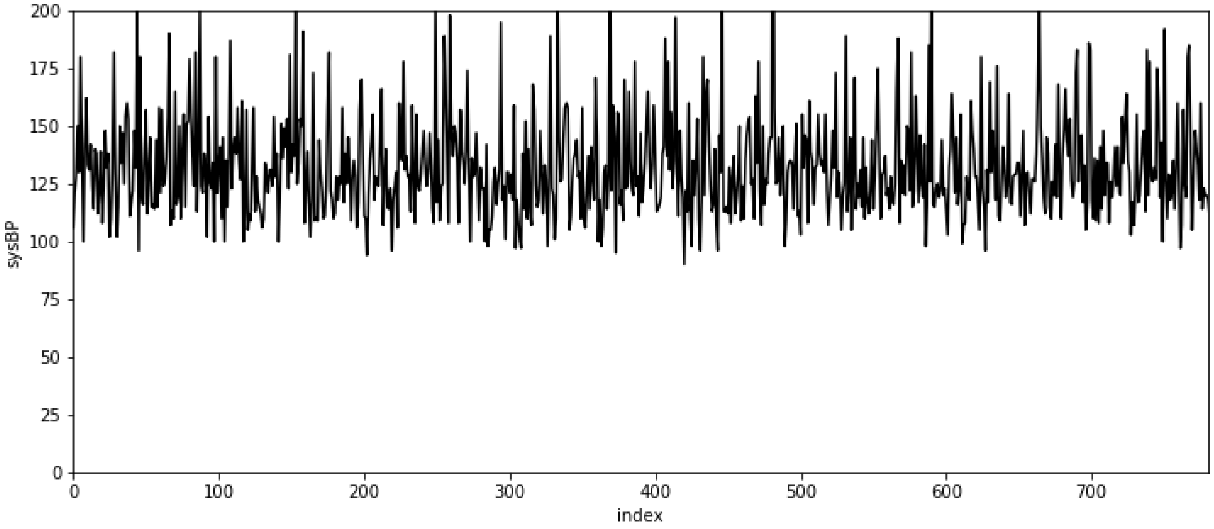
In [19]:
```
sns.heatmap(df.corr(), annot=True)
```

Out[19]:
```
<AxesSubplot:>
```



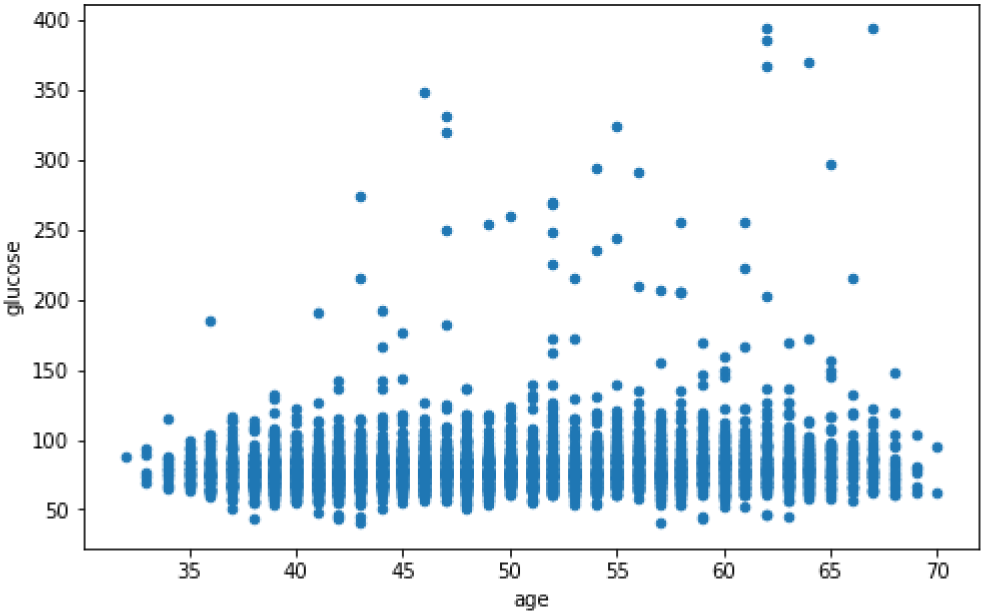
In [20]:
```
df['sysBP'].plot(figsize=(12, 5), color='black') # color and figsize changed

plt.xlim(0, 780) # range for x-axis
plt.ylim(0, 200) # range for x-axis
plt.xlabel('index')
plt.ylabel('sysBP')
```

Out[20]:
```
Text(0, 0.5, 'sysBP')
```

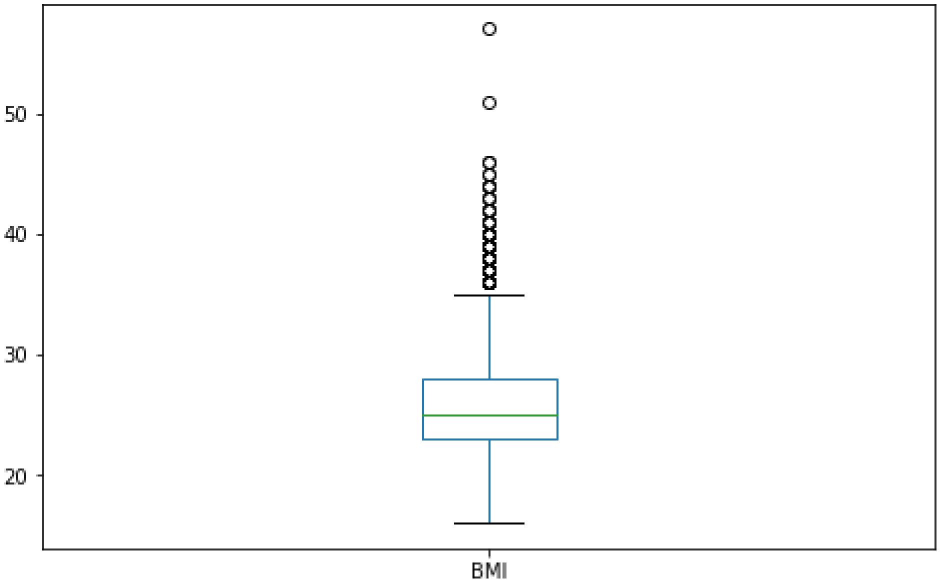


In [21]:
```
df.plot.scatter('age', 'glucose', figsize=(8, 5))
```

Out[21]:
```
<AxesSubplot:xlabel='age', ylabel='glucose'>
```

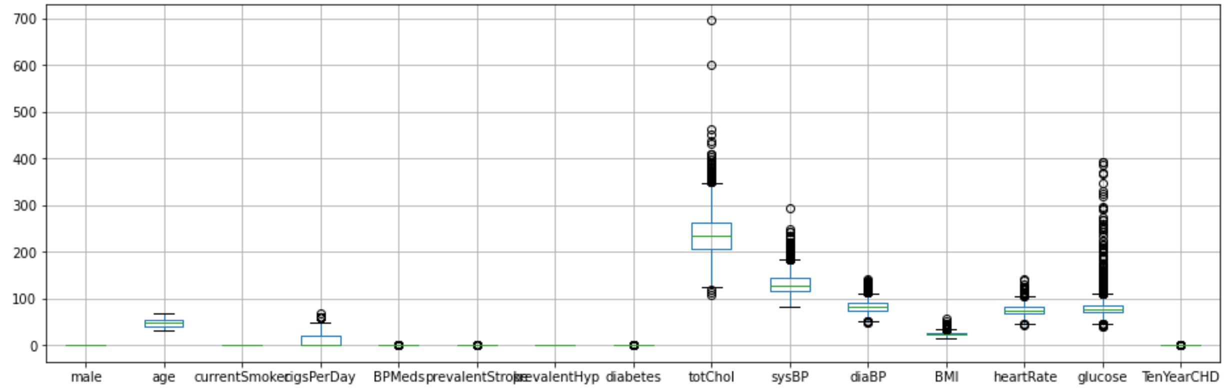


In [22]:
```
df['BMI'].plot.box(figsize=(8, 5));
```

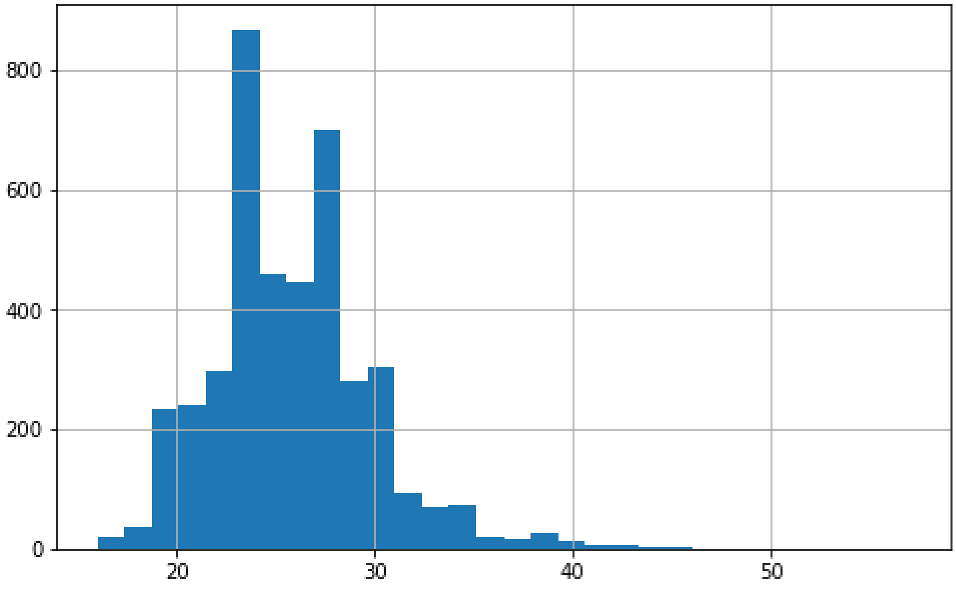


In [23]:
```
df.boxplot(figsize=(16, 5)) # or df.plot.box()
```

Out[23]:
```
<AxesSubplot:>
```

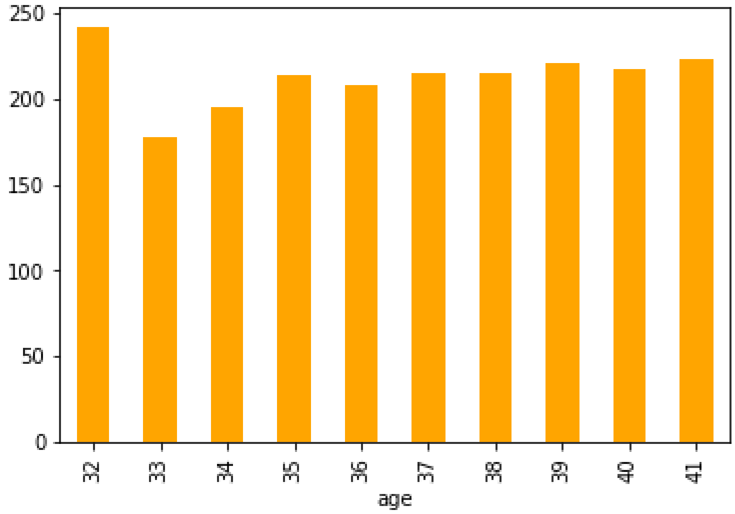


In [24]:
```
df['BMI'].hist(bins=30, figsize=(8, 5)); # we can specify the number of bins
```



In [25]:

```
df_avg_BP = df.groupby('age')['totChol'].mean()
df_avg_BP[:10].plot.bar(color='orange');
```



In [ ]:

```
```