cleaning and preprocessing

importing libraries

In [1]:
```
import numpy as np
import pandas as pd
```

importing dataset

In [3]:
```
data=pd.read_csv(r"C:\Users\user\Downloads\8_BreastCancerPrediction - 8_BreastCar
data
```

Out[3]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_m
0	842302	M	17.99	10.38	122.80	1001.0	0.11
1	842517	M	20.57	17.77	132.90	1326.0	0.08
2	84300903	M	19.69	21.25	130.00	1203.0	0.10
3	84348301	M	11.42	20.38	77.58	386.1	0.14
4	84358402	M	20.29	14.34	135.10	1297.0	0.10
...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.11
565	926682	M	20.13	28.25	131.20	1261.0	0.09
566	926954	M	16.60	28.08	108.30	858.1	0.08
567	927241	M	20.60	29.33	140.10	1265.0	0.11
568	92751	B	7.76	24.54	47.92	181.0	0.05

569 rows × 32 columns

To print initial rows

In [4]:
```
data.head()
```

Out[4]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_m
0	842302	M	17.99	10.38	122.80	1001.0	0.11
1	842517	M	20.57	17.77	132.90	1326.0	0.08
2	84300903	M	19.69	21.25	130.00	1203.0	0.10
3	84348301	M	11.42	20.38	77.58	386.1	0.14
4	84358402	M	20.29	14.34	135.10	1297.0	0.10

5 rows × 32 columns

To print last rows

In [5]:
```
data.tail()
```

Out[5]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mea
564	926424	M	21.56	22.39	142.00	1479.0	0.1110
565	926682	M	20.13	28.25	131.20	1261.0	0.0978
566	926954	M	16.60	28.08	108.30	858.1	0.0845
567	927241	M	20.60	29.33	140.10	1265.0	0.1178
568	92751	B	7.76	24.54	47.92	181.0	0.0526

5 rows × 32 columns

describing the data

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

Out[6]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400

8 rows × 31 columns

To print shape and size

In [7]:
```
print(np.shape(data))
```

```
(569, 32)
```

In [8]:
```
print(np.size(data))
```

```
18208
```

finding missing values

In [9]:
```
data.isnull()
```

Out[9]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
564	False	False	False	False	False	False	False
565	False	False	False	False	False	False	False
566	False	False	False	False	False	False	False
567	False	False	False	False	False	False	False
568	False	False	False	False	False	False	False

569 rows × 32 columns

In [ ]:
```
filling missing values
```

In [10]:
```
data.dropna(0)
```

Out[10]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness
0	842302	M	17.99	10.38	122.80	1001.0	0
1	842517	M	20.57	17.77	132.90	1326.0	0
2	84300903	M	19.69	21.25	130.00	1203.0	0
3	84348301	M	11.42	20.38	77.58	386.1	0
4	84358402	M	20.29	14.34	135.10	1297.0	0
...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0
565	926682	M	20.13	28.25	131.20	1261.0	0
566	926954	M	16.60	28.08	108.30	858.1	0
567	927241	M	20.60	29.33	140.10	1265.0	0
568	92751	B	7.76	24.54	47.92	181.0	0

Visualization

In [12]:

```
data=data[["radius_mean","texture_mean"]]
data
```

Out[12]:

	radius_mean	texture_mean
0	17.99	10.38
1	20.57	17.77
2	19.69	21.25
3	11.42	20.38
4	20.29	14.34
...	...	...
564	21.56	22.39
565	20.13	28.25
566	16.60	28.08
567	20.60	29.33
568	7.76	24.54

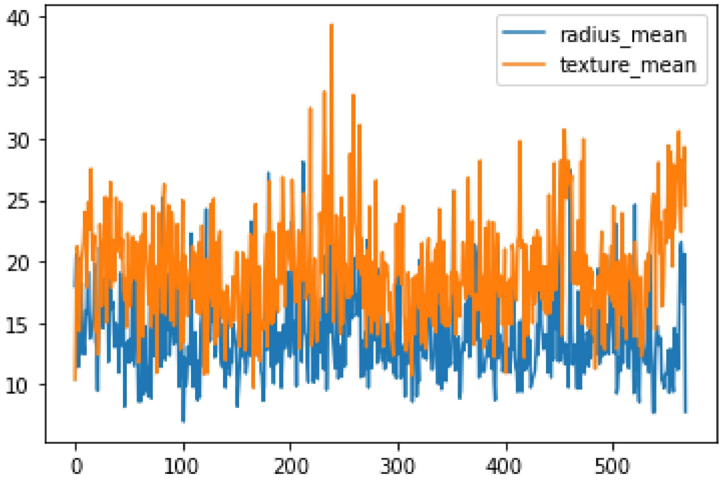
569 rows × 2 columns

In [13]:

```
data.plot.line()
```

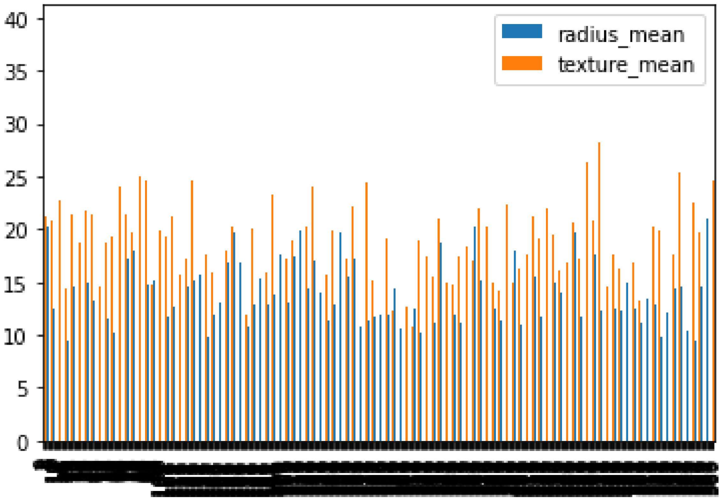
Out[13]:

```
<AxesSubplot:>
```



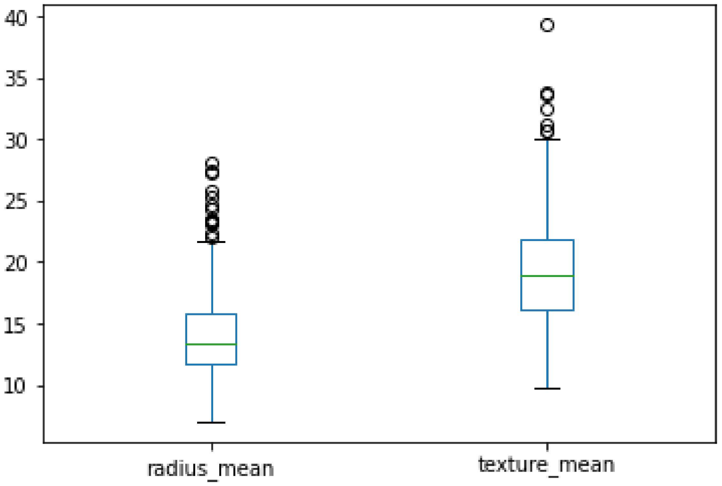
In [14]:
```
data.plot.bar()
```

Out[14]:
```
<AxesSubplot:>
```



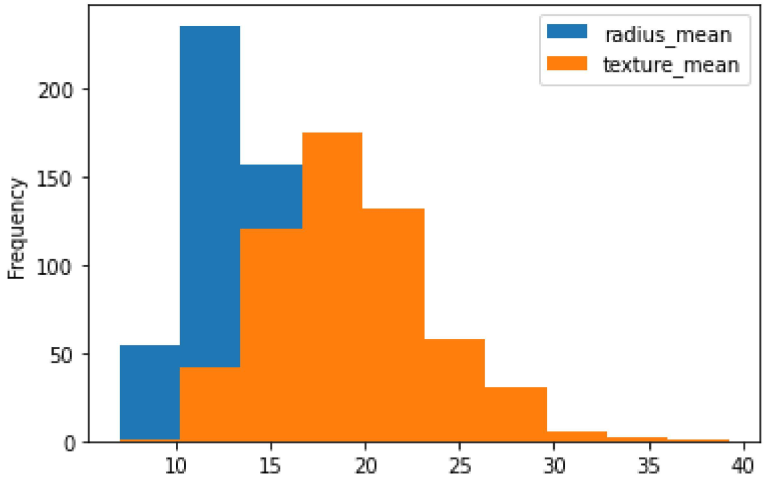
In [15]:
```
data.plot.box()
```

Out[15]:
```
<AxesSubplot:>
```



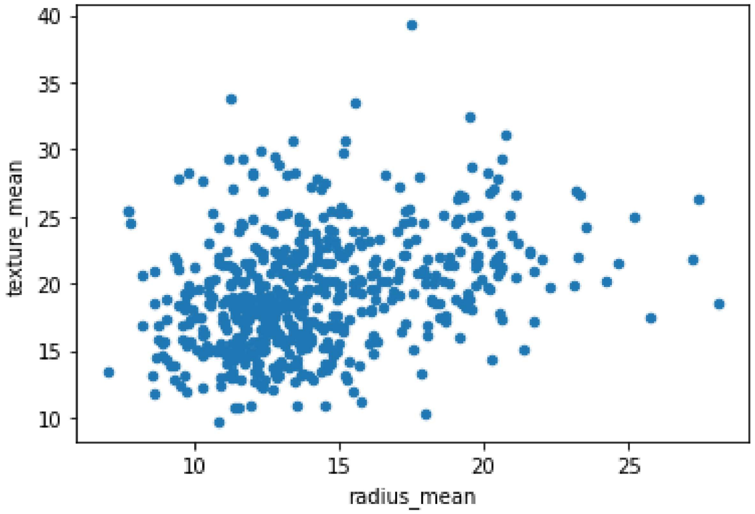
```
In [16]: data.plot.hist()
```

```
Out[16]: <AxesSubplot:ylabel='Frequency'>
```



```
In [17]: data.plot.scatter(x="radius_mean",y="texture_mean")
```

```
Out[17]: <AxesSubplot:xlabel='radius_mean', ylabel='texture_mean'>
```



In [18]:

```
data.plot.pie(subplots=True)
```

Out[18]:

```
array([<AxesSubplot:ylabel='radius_mean'>,
       <AxesSubplot:ylabel='texture_mean'>], dtype=object)
```

In [ ]: