data clean,preprocess and visualization

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

Import dataset

In [25]:
```
data=pd.read_csv(r"C:\Users\user\Downloads\8_BreastCancerPrediction.csv")
```

print data

In [26]:
```
data
```

Out[26]:

	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	0.0
2	84300903	M	19.69	21.25	130.00	1203.0	0.1
3	84348301	M	11.42	20.38	77.58	386.1	0.1
4	84358402	M	20.29	14.34	135.10	1297.0	0.1
...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.
565	926682	M	20.13	28.25	131.20	1261.0	0.0
566	926954	M	16.60	28.08	108.30	858.1	0.0
567	927241	M	20.60	29.33	140.10	1265.0	0.
568	92751	B	7.76	24.54	47.92	181.0	0.0

569 rows × 33 columns

print first 10 rows using head

In [27]:
```
data.head(10)
```

Out[27]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_me
0	842302	M	17.99	10.38	122.80	1001.0	0.118
1	842517	M	20.57	17.77	132.90	1326.0	0.084
2	84300903	M	19.69	21.25	130.00	1203.0	0.109
3	84348301	M	11.42	20.38	77.58	386.1	0.142
4	84358402	M	20.29	14.34	135.10	1297.0	0.100
5	843786	M	12.45	15.70	82.57	477.1	0.127
6	844359	M	18.25	19.98	119.60	1040.0	0.094
7	84458202	M	13.71	20.83	90.20	577.9	0.118
8	844981	M	13.00	21.82	87.50	519.8	0.127
9	84501001	M	12.46	24.04	83.97	475.9	0.118

10 rows × 33 columns

print last 10 rows using tail

In [28]:
```
data.tail(5)
```

Out[28]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_me
564	926424	M	21.56	22.39	142.00	1479.0	0.111
565	926682	M	20.13	28.25	131.20	1261.0	0.097
566	926954	M	16.60	28.08	108.30	858.1	0.084
567	927241	M	20.60	29.33	140.10	1265.0	0.117
568	92751	B	7.76	24.54	47.92	181.0	0.052

5 rows × 33 columns

print describe of dataset

In [29]:
```
data.describe()
```

Out[29]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mea
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.00000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.09636
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.01406
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.05263
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.08637
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.09587
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.10530
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.16340

8 rows × 32 columns

Number elements in dataset

In [30]:
```
data.size
```

Out[30]:
```
18777
```

print shape of dataset

In [31]:
```
data.shape
```

Out[31]:
```
(569, 33)
```

print empty or not

In [32]:
```
data.isna()
```

Out[32]:

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

569 rows × 33 columns

In [33]:
```
data.isnull().sum()
```

Out[33]:
```
id                           0
diagnosis                    0
radius_mean                  0
texture_mean                 0
perimeter_mean               0
area_mean                    0
smoothness_mean              0
compactness_mean             0
concavity_mean               0
concave points_mean          0
symmetry_mean                0
fractal_dimension_mean       0
radius_se                    0
texture_se                   0
perimeter_se                 0
area_se                      0
smoothness_se                0
compactness_se               0
concavity_se                 0
concave points_se            0
symmetry_se                  0
fractal_dimension_se         0
radius_worst                 0
texture_worst                0
perimeter_worst              0
area_worst                   0
smoothness_worst             0
compactness_worst            0
concavity_worst              0
concave points_worst         0
symmetry_worst               0
fractal_dimension_worst      0
Unnamed: 32                569
dtype: int64
```

In [34]:

```
data1 = data.fillna(value=10)
data1
```

Out[34]:

	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.

569 rows × 33 columns

In [36]:

```
data1 = data[["id","radius_mean"]]
data1
```

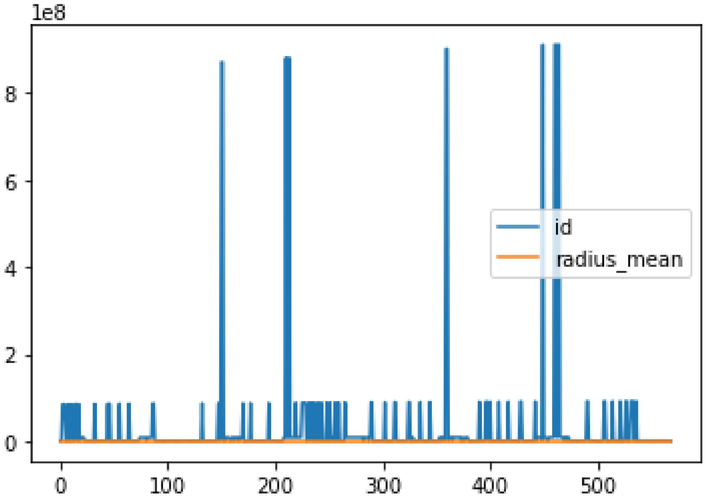
Out[36]:

	id	radius_mean
0	842302	17.99
1	842517	20.57
2	84300903	19.69
3	84348301	11.42
4	84358402	20.29
...	...	...
564	926424	21.56
565	926682	20.13
566	926954	16.60
567	927241	20.60
568	92751	7.76

569 rows × 2 columns

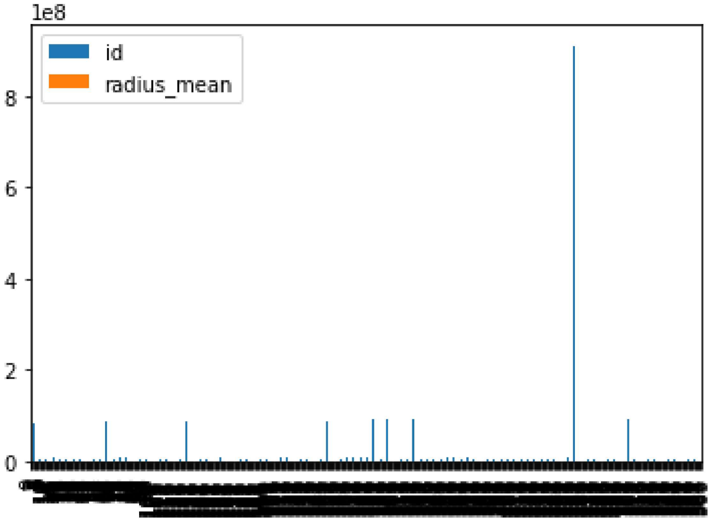
In [37]:
```
data1.plot.line()
```

Out[37]:
```
<AxesSubplot:>
```



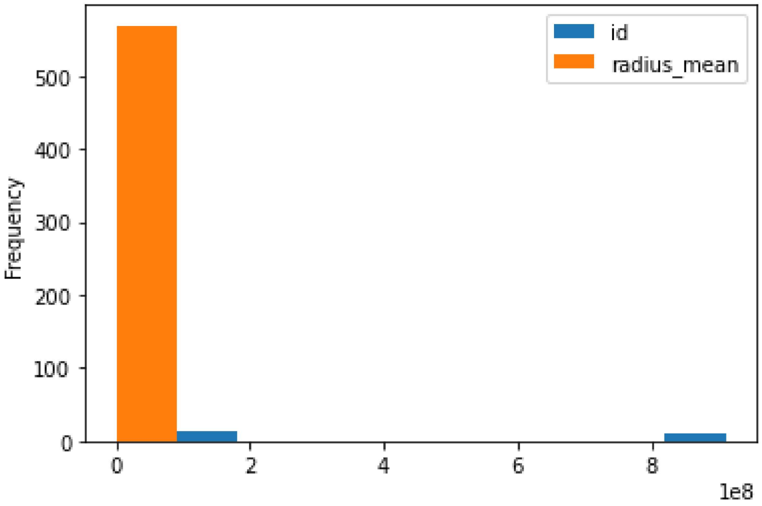
In [38]:
```
data1.plot.bar()
```

Out[38]:
```
<AxesSubplot:>
```



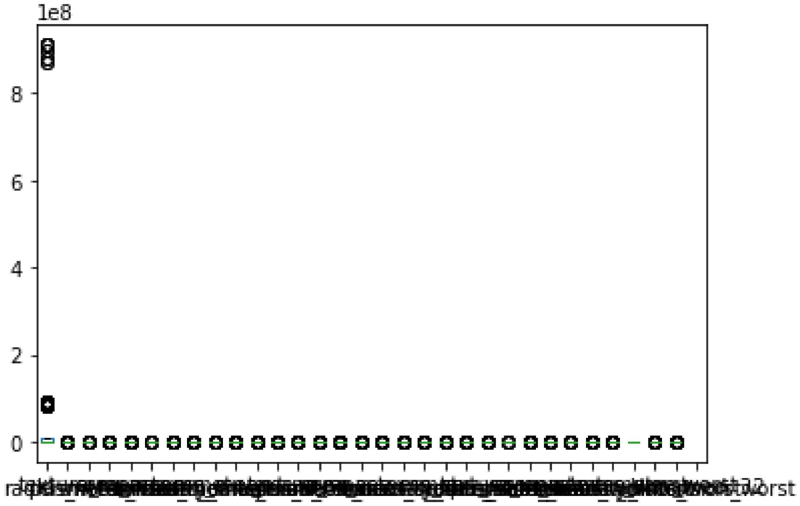
In [39]:
```
data1.plot.hist()
```

Out[39]:
```
<AxesSubplot:ylabel='Frequency'>
```



In [46]:
```
data.plot.box("id")
```

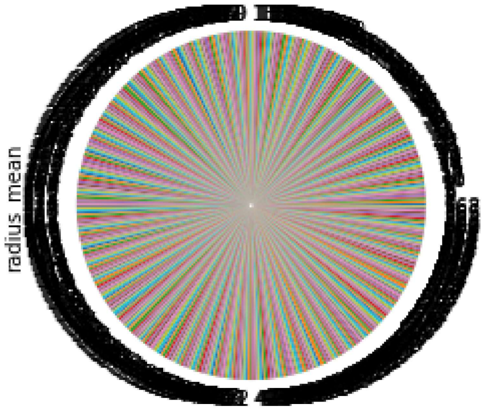
Out[46]:
```
<AxesSubplot:>
```



In [42]:
```
data2 = data1["radius_mean"]
```

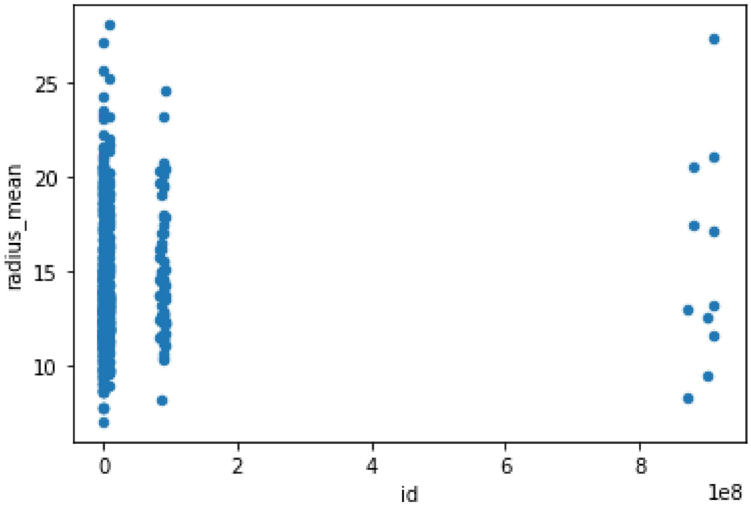
```
In [43]: data2.plot.pie()
```

```
Out[43]: <AxesSubplot:ylabel='radius_mean'>
```



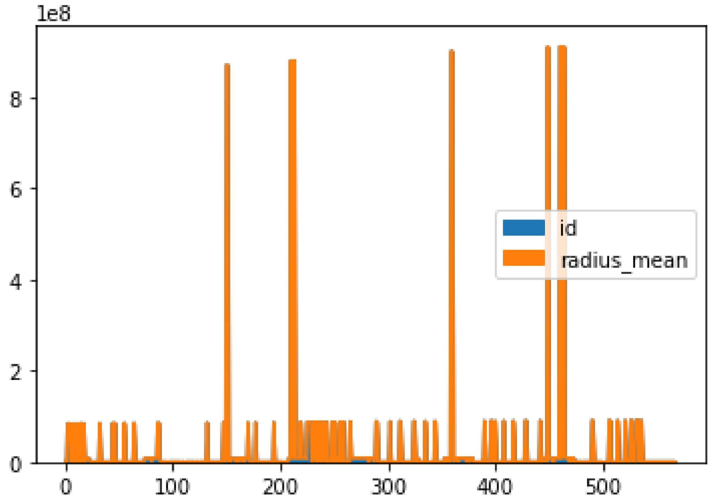
```
In [47]: data1.plot.scatter("id","radius_mean")
```

```
Out[47]: <AxesSubplot:xlabel='id', ylabel='radius_mean'>
```



In [45]:
```
data1.plot.area()
```

Out[45]:
```
<AxesSubplot:>
```



In [ ]: