In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

With the two datasets given (refer to drive) - Frame a problem statement, clean, preprocess and visulaize the data and interpret your conclusion

In [2]:

```
df=pd.read_csv("pre.csv")
df
```

Out[2]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	c
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	
...	...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	
566	926954	M	16.60	28.08	108.30	858.1	0.08455	
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	
568	92751	B	7.76	24.54	47.92	181.0	0.05263	

569 rows × 33 columns

In [3]:

```
df.head(50)
```

Out[3]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cc
0	842302	M	17.990	10.38	122.80	1001.0	0.11840	
1	842517	M	20.570	17.77	132.90	1326.0	0.08474	
2	84300903	M	19.690	21.25	130.00	1203.0	0.10960	
3	84348301	M	11.420	20.38	77.58	386.1	0.14250	
4	84358402	M	20.290	14.34	135.10	1297.0	0.10030	

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cc
5	843786	M	12.450	15.70	82.57	477.1	0.12780	
6	844359	M	18.250	19.98	119.60	1040.0	0.09463	
7	84458202	M	13.710	20.83	90.20	577.9	0.11890	
8	844981	M	13.000	21.82	87.50	519.8	0.12730	
9	84501001	M	12.460	24.04	83.97	475.9	0.11860	
10	845636	M	16.020	23.24	102.70	797.8	0.08206	
11	84610002	M	15.780	17.89	103.60	781.0	0.09710	
12	846226	M	19.170	24.80	132.40	1123.0	0.09740	
13	846381	M	15.850	23.95	103.70	782.7	0.08401	
14	84667401	M	13.730	22.61	93.60	578.3	0.11310	
15	84799002	M	14.540	27.54	96.73	658.8	0.11390	
16	848406	M	14.680	20.13	94.74	684.5	0.09867	
17	84862001	M	16.130	20.68	108.10	798.8	0.11700	
18	849014	M	19.810	22.15	130.00	1260.0	0.09831	
19	8510426	B	13.540	14.36	87.46	566.3	0.09779	
20	8510653	B	13.080	15.71	85.63	520.0	0.10750	
21	8510824	B	9.504	12.44	60.34	273.9	0.10240	
22	8511133	M	15.340	14.26	102.50	704.4	0.10730	
23	851509	M	21.160	23.04	137.20	1404.0	0.09428	
24	852552	M	16.650	21.38	110.00	904.6	0.11210	
25	852631	M	17.140	16.40	116.00	912.7	0.11860	
26	852763	M	14.580	21.53	97.41	644.8	0.10540	
27	852781	M	18.610	20.25	122.10	1094.0	0.09440	
28	852973	M	15.300	25.27	102.40	732.4	0.10820	
29	853201	M	17.570	15.05	115.00	955.1	0.09847	
30	853401	M	18.630	25.11	124.80	1088.0	0.10640	
31	853612	M	11.840	18.70	77.93	440.6	0.11090	
32	85382601	M	17.020	23.98	112.80	899.3	0.11970	
33	854002	M	19.270	26.47	127.90	1162.0	0.09401	
34	854039	M	16.130	17.88	107.00	807.2	0.10400	
35	854253	M	16.740	21.59	110.10	869.5	0.09610	
36	854268	M	14.250	21.72	93.63	633.0	0.09823	
37	854941	B	13.030	18.42	82.61	523.8	0.08983	

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cc
38	855133	M	14.990	25.20	95.54	698.8	0.09387	
39	855138	M	13.480	20.82	88.40	559.2	0.10160	
40	855167	M	13.440	21.58	86.18	563.0	0.08162	
41	855563	M	10.950	21.35	71.90	371.1	0.12270	
42	855625	M	19.070	24.81	128.30	1104.0	0.09081	
43	856106	M	13.280	20.28	87.32	545.2	0.10410	
44	85638502	M	13.170	21.81	85.42	531.5	0.09714	
45	857010	M	18.650	17.60	123.70	1076.0	0.10990	
46	85713702	B	8.196	16.84	51.71	201.9	0.08600	
47	85715	M	13.170	18.66	85.98	534.6	0.11580	
48	857155	B	12.050	14.63	78.04	449.3	0.10310	
49	857156	B	13.490	22.30	86.91	561.0	0.08752	

50 rows × 33 columns

In [4]:

```
df.tail(50)
```

Out[4]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	c
519	917080	B	12.750	16.70	82.51	493.8	0.11250	
520	917092	B	9.295	13.90	59.96	257.8	0.13710	
521	91762702	M	24.630	21.60	165.50	1841.0	0.10300	
522	91789	B	11.260	19.83	71.30	388.1	0.08511	
523	917896	B	13.710	18.68	88.73	571.0	0.09916	
524	917897	B	9.847	15.68	63.00	293.2	0.09492	
525	91805	B	8.571	13.10	54.53	221.3	0.10360	
526	91813701	B	13.460	18.75	87.44	551.1	0.10750	
527	91813702	B	12.340	12.27	78.94	468.5	0.09003	
528	918192	B	13.940	13.17	90.31	594.2	0.12480	
529	918465	B	12.070	13.44	77.83	445.2	0.11000	
530	91858	B	11.750	17.56	75.89	422.9	0.10730	
531	91903901	B	11.670	20.02	75.21	416.2	0.10160	
532	91903902	B	13.680	16.33	87.76	575.5	0.09277	
533	91930402	M	20.470	20.67	134.70	1299.0	0.09156	

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
534	919537	B	10.960	17.62	70.79	365.6	0.09687
535	919555	M	20.550	20.86	137.80	1308.0	0.10460
536	91979701	M	14.270	22.55	93.77	629.8	0.10380
537	919812	B	11.690	24.44	76.37	406.4	0.12360
538	921092	B	7.729	25.49	47.98	178.8	0.08098
539	921362	B	7.691	25.44	48.34	170.4	0.08668
540	921385	B	11.540	14.44	74.65	402.9	0.09984
541	921386	B	14.470	24.99	95.81	656.4	0.08837
542	921644	B	14.740	25.42	94.70	668.6	0.08275
543	922296	B	13.210	28.06	84.88	538.4	0.08671
544	922297	B	13.870	20.70	89.77	584.8	0.09578
545	922576	B	13.620	23.23	87.19	573.2	0.09246
546	922577	B	10.320	16.35	65.31	324.9	0.09434
547	922840	B	10.260	16.58	65.85	320.8	0.08877
548	923169	B	9.683	19.34	61.05	285.7	0.08491
549	923465	B	10.820	24.21	68.89	361.6	0.08192
550	923748	B	10.860	21.48	68.51	360.5	0.07431
551	923780	B	11.130	22.44	71.49	378.4	0.09566
552	924084	B	12.770	29.43	81.35	507.9	0.08276
553	924342	B	9.333	21.94	59.01	264.0	0.09240
554	924632	B	12.880	28.92	82.50	514.3	0.08123
555	924934	B	10.290	27.61	65.67	321.4	0.09030
556	924964	B	10.160	19.59	64.73	311.7	0.10030
557	925236	B	9.423	27.88	59.26	271.3	0.08123
558	925277	B	14.590	22.68	96.39	657.1	0.08473
559	925291	B	11.510	23.93	74.52	403.5	0.09261
560	925292	B	14.050	27.15	91.38	600.4	0.09929
561	925311	B	11.200	29.37	70.67	386.0	0.07449
562	925622	M	15.220	30.62	103.40	716.9	0.10480
563	926125	M	20.920	25.09	143.00	1347.0	0.10990
564	926424	M	21.560	22.39	142.00	1479.0	0.11100
565	926682	M	20.130	28.25	131.20	1261.0	0.09780
566	926954	M	16.600	28.08	108.30	858.1	0.08455

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	c
567	927241	M	20.600	29.33	140.10	1265.0	0.11780	
568	92751	B	7.760	24.54	47.92	181.0	0.05263	

50 rows × 33 columns

In [5]:
```
df.describe()
```

Out[5]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	comp
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 × 32 columns

In [6]:
```
df.shape
```

Out[6]:
```
(569, 33)
```

In [7]:
```
df.size
```

Out[7]:
```
18777
```

In [8]:
```
c=df.head(20)
c
```

Out[8]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cc
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cc
5	843786	M	12.45	15.70	82.57	477.1	0.12780	
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	
7	84458202	M	13.71	20.83	90.20	577.9	0.11890	
8	844981	M	13.00	21.82	87.50	519.8	0.12730	
9	84501001	M	12.46	24.04	83.97	475.9	0.11860	
10	845636	M	16.02	23.24	102.70	797.8	0.08206	
11	84610002	M	15.78	17.89	103.60	781.0	0.09710	
12	846226	M	19.17	24.80	132.40	1123.0	0.09740	
13	846381	M	15.85	23.95	103.70	782.7	0.08401	
14	84667401	M	13.73	22.61	93.60	578.3	0.11310	
15	84799002	M	14.54	27.54	96.73	658.8	0.11390	
16	848406	M	14.68	20.13	94.74	684.5	0.09867	
17	84862001	M	16.13	20.68	108.10	798.8	0.11700	
18	849014	M	19.81	22.15	130.00	1260.0	0.09831	
19	8510426	B	13.54	14.36	87.46	566.3	0.09779	

20 rows × 33 columns

In [9]:

```
df.fillna(value=0)
```

Out[9]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	c
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	
...	...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	
566	926954	M	16.60	28.08	108.30	858.1	0.08455	
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	
568	92751	B	7.76	24.54	47.92	181.0	0.05263	

569 rows × 33 columns

In [10]:

```
df.isna()
```

Out[10]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	comp
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	

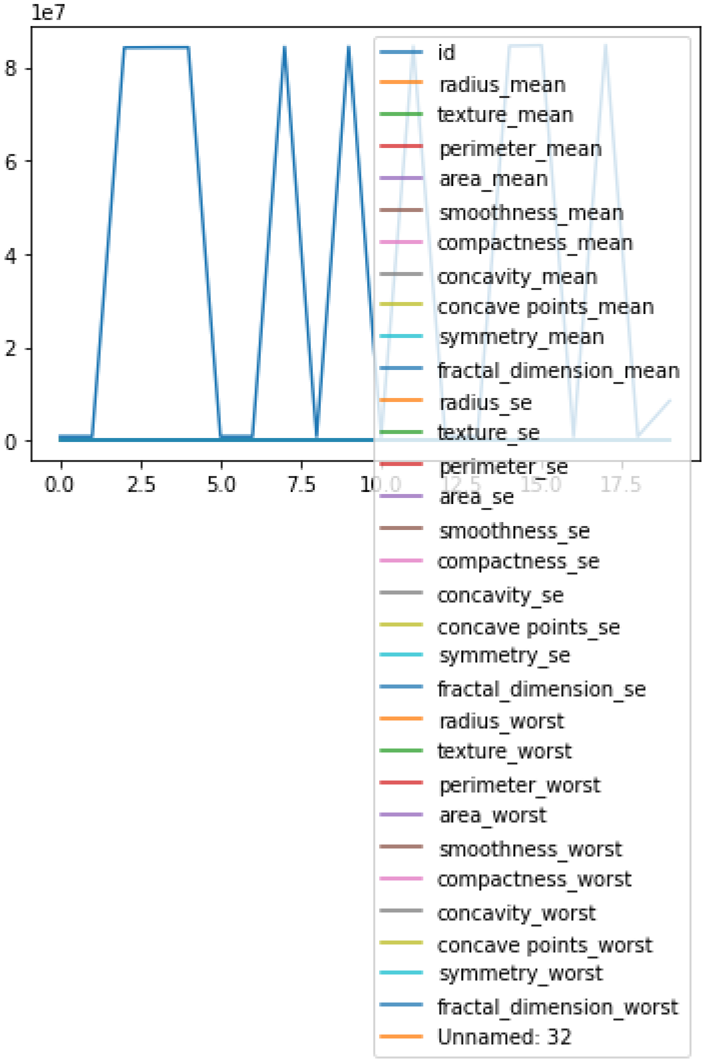
569 rows × 33 columns

In [12]:

```
c.plot()
```

Out[12]:

```
<AxesSubplot:>
```

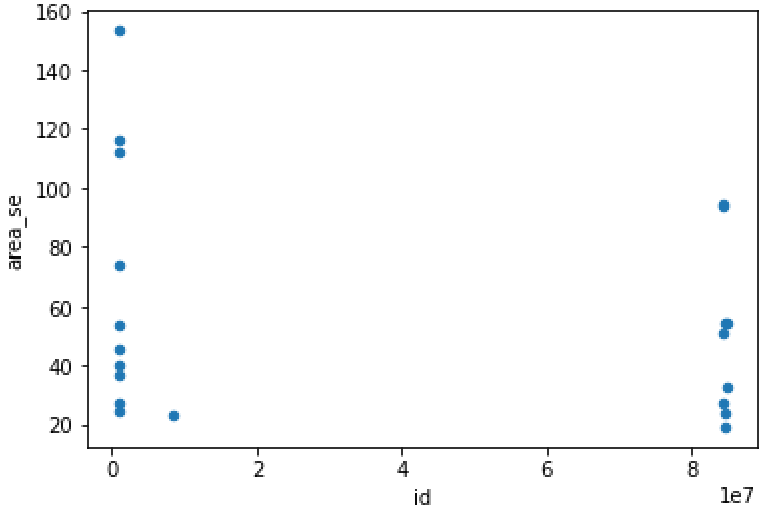


In [13]:

```
c.plot.scatter(x="id",y="area_se")
```

Out[13]:

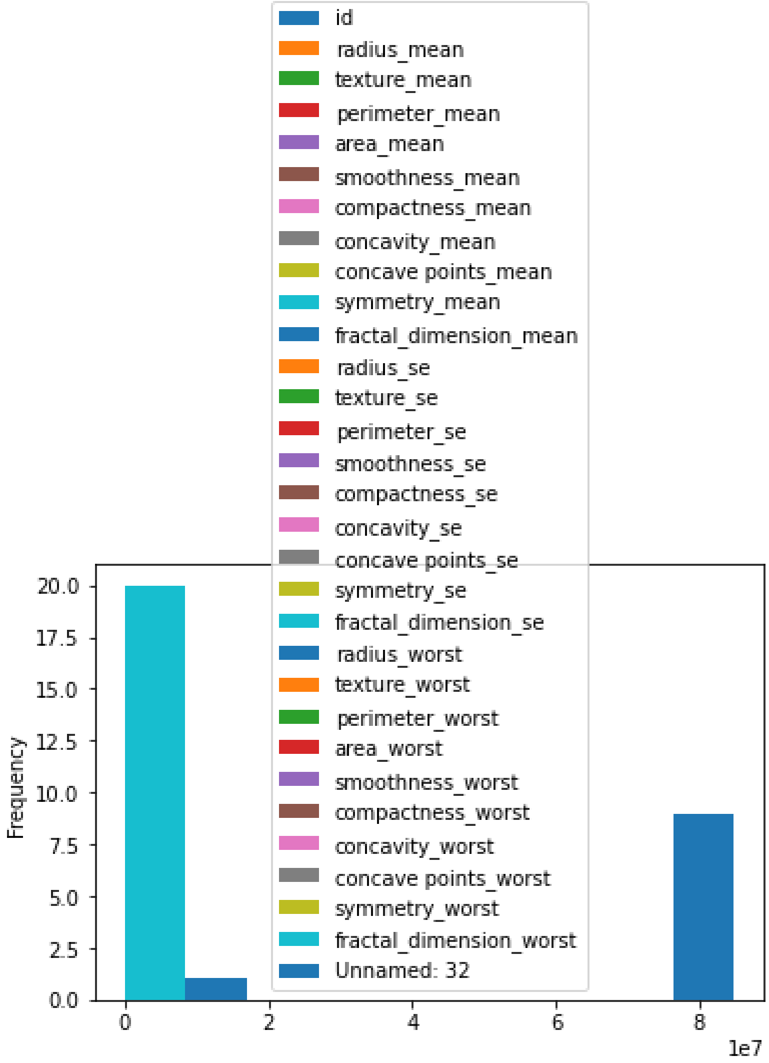
```
<AxesSubplot:xlabel='id', ylabel='area_se'>
```



In [14]:

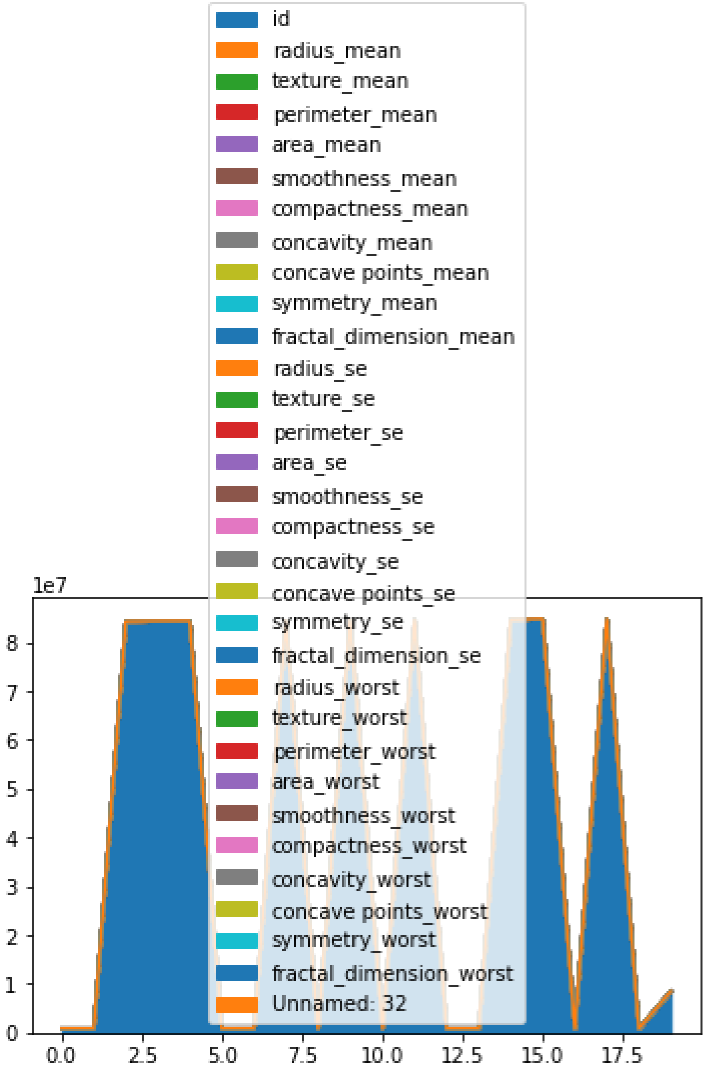
```
c.plot.hist(x="area_se")
```

```
Out[14]: <AxesSubplot:ylabel='Frequency'>
```



```
In [15]: c.plot.area()
```

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

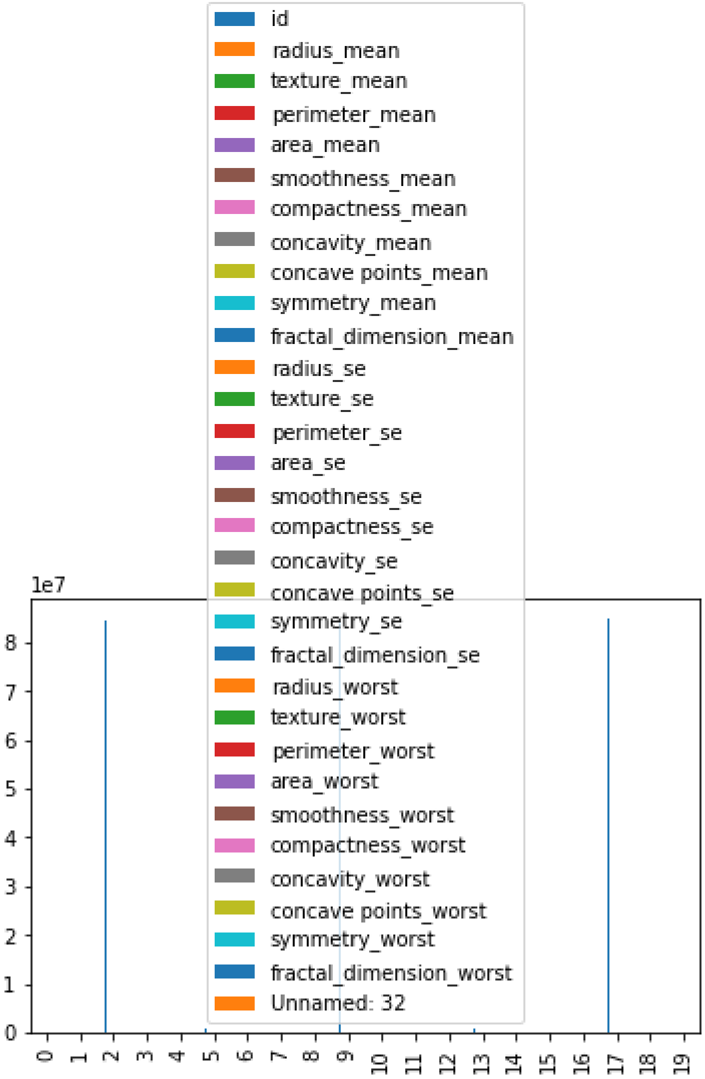


In [16]:

```
c.plot.bar()
```

Out[16]:

```
<AxesSubplot:>
```

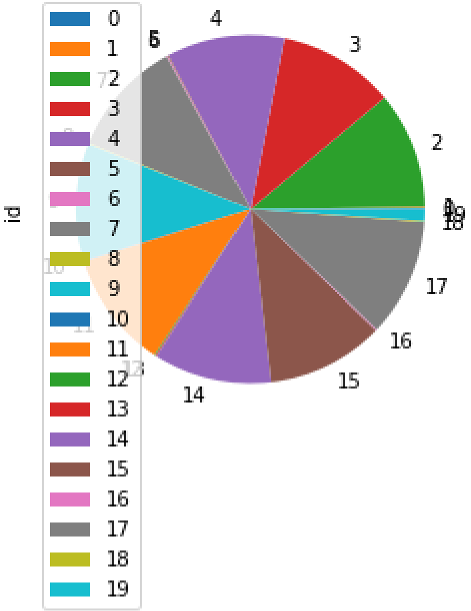


In [17]:

```
c.plot.pie(y="id")
```

Out[17]:

```
<AxesSubplot:ylabel='id'>
```



In [ ]: