

# EDA-Exploitory Data Analysis -Breast Cancer Dataset

## import libraries

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

## import dataset

```
In [2]: data=pd.read_csv(r"E:\154\8_BreastCancerPrediction - 8_BreastCancerPrediction.csv")
```

```
In [3]: display(data)
```

0	842302	M	17.99	10.38	122.80	1001.0	0.1
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	0.1
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	0.1
568	92751	B	7.76	24.54	47.92	181.0	0.0
569 rows × 32 columns							

## To display top 10 rows

In [4]: `data.head(10)`

Out[4]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
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
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 rows × 32 columns

In [ ]:

## To display last 5 rows

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

Out[5]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
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

5 rows × 32 columns

```
In [6]: data.dtypes
```

```
Out[6]: id                  int64
diagnosis          object
radius_mean        float64
texture_mean       float64
perimeter_mean    float64
area_mean          float64
smoothness_mean   float64
compactness_mean  float64
concavity_mean    float64
concave_points_mean float64
symmetry_mean     float64
fractal_dimension_mean float64
radius_se          float64
texture_se         float64
perimeter_se       float64
area_se            float64
smoothness_se     float64
compactness_se    float64
concavity_se      float64
concave_points_se float64
symmetry_se       float64
fractal_dimension_se float64
radius_worst       float64
texture_worst      float64
perimeter_worst   float64
area_worst         float64
smoothness_worst  float64
compactness_worst  float64
concavity_worst   float64
concave_points_worst float64
symmetry_worst    float64
fractal_dimension_worst float64
dtype: object
```

## To view statistical summary

In [7]: `data.describe()`

Out[7]:

	<code>id</code>	<code>radius_mean</code>	<code>texture_mean</code>	<code>perimeter_mean</code>	<code>area_mean</code>	<code>smoothness_mean</code>	<code>compactness_mean</code>	<code>concavity_mean</code>	<code>concave points_mean</code>	<code>symmetry_mean</code>	<code>fractal_dimension_mean</code>	<code>radius_se</code>	<code>texture_se</code>	<code>perimeter_se</code>	<code>area_se</code>	<code>smoothness_se</code>	<code>compactness_se</code>	<code>concavity_se</code>	<code>concave points_se</code>	<code>symmetry_se</code>	<code>fractal_dimension_se</code>	<code>radius_worst</code>	<code>texture_worst</code>	<code>perimeter_worst</code>	<code>area_worst</code>	<code>smoothness_worst</code>	<code>compactness_worst</code>	<code>concavity_worst</code>	<code>concave points_worst</code>	<code>symmetry_worst</code>	<code>fractal_dimension_worst</code>
<code>count</code>	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000			
<code>mean</code>	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360	0.096360		
<code>std</code>	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064	0.014064		
<code>min</code>	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	0.052630	
<code>25%</code>	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370	0.086370		
<code>50%</code>	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870	0.095870		
<code>75%</code>	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300	0.105300		
<code>max</code>	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	0.163400	

8 rows × 31 columns

## To Print no of elements

In [8]: `data.size`

Out[8]: 18208

In [9]: `data.ndim`

Out[9]: 2

## To print no of rows and columns

In [10]: `data.shape`

Out[10]: (569, 32)

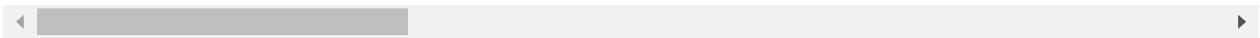
## To find missing values

In [11]: `data.isna()`

Out[11]:

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

569 rows × 32 columns



## To drop null values with constatns

In [12]: `data.fillna(5)`

Out[12]:

	<code>id</code>	<code>diagnosis</code>	<code>radius_mean</code>	<code>texture_mean</code>	<code>perimeter_mean</code>	<code>area_mean</code>	<code>smoothness_mean</code>
0	842302	M	17.99	10.38	122.80	1001.0	0.1184
1	842517	M	20.57	17.77	132.90	1326.0	0.0847
2	84300903	M	19.69	21.25	130.00	1203.0	0.1096
3	84348301	M	11.42	20.38	77.58	386.1	0.1425
4	84358402	M	20.29	14.34	135.10	1297.0	0.1003
...	...	...	...	...	...	...	...
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

569 rows × 32 columns

In [13]: `data.dropna()`

Out[13]:

	<code>id</code>	<code>diagnosis</code>	<code>radius_mean</code>	<code>texture_mean</code>	<code>perimeter_mean</code>	<code>area_mean</code>	<code>smoothness_mean</code>
0	842302	M	17.99	10.38	122.80	1001.0	0.1184
1	842517	M	20.57	17.77	132.90	1326.0	0.0847
2	84300903	M	19.69	21.25	130.00	1203.0	0.1096
3	84348301	M	11.42	20.38	77.58	386.1	0.1425
4	84358402	M	20.29	14.34	135.10	1297.0	0.1003
...	...	...	...	...	...	...	...
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

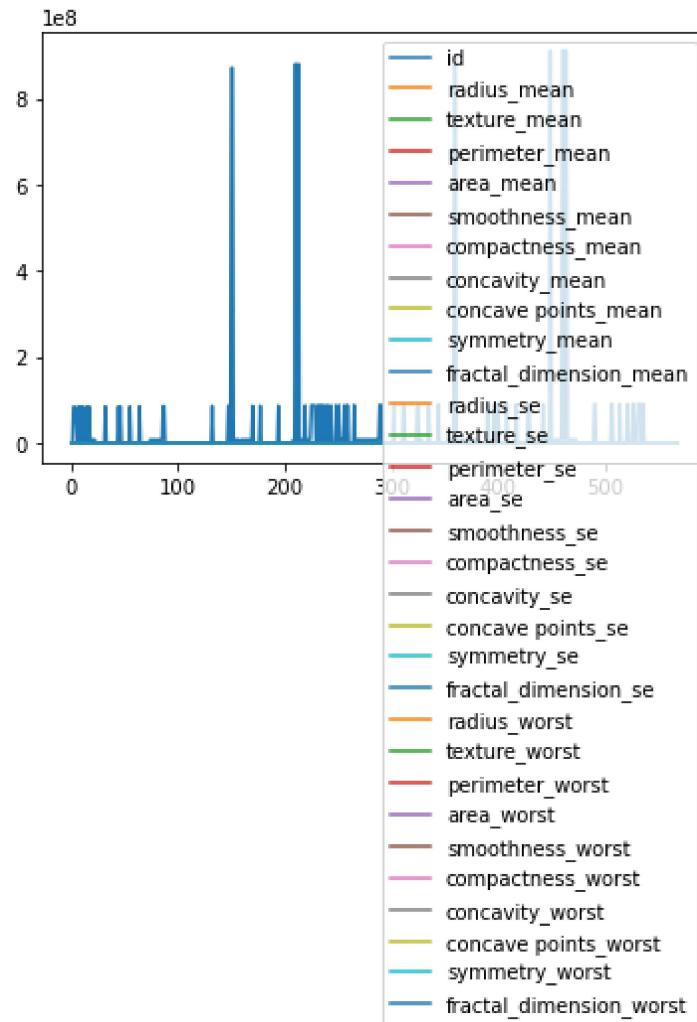
569 rows × 32 columns

## Line Plot

Type *Markdown* and *LaTeX*:  $\alpha^2$

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

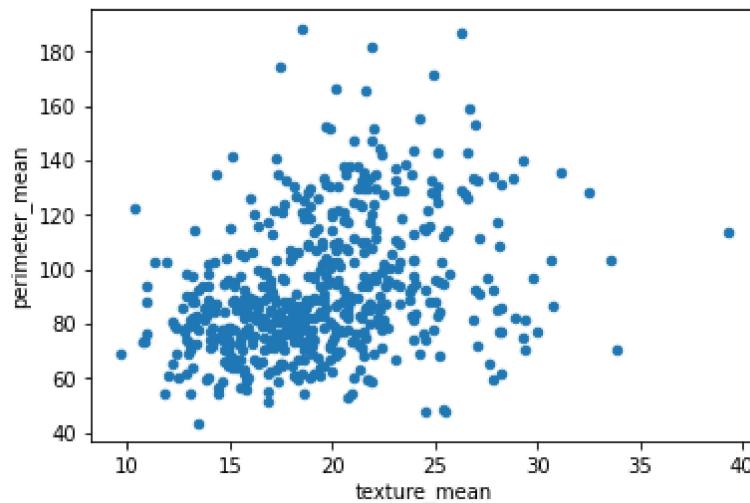
Out[14]: <AxesSubplot:>



## Scatter Plot

```
In [15]: data.plot.scatter(x='texture_mean',y='perimeter_mean')
```

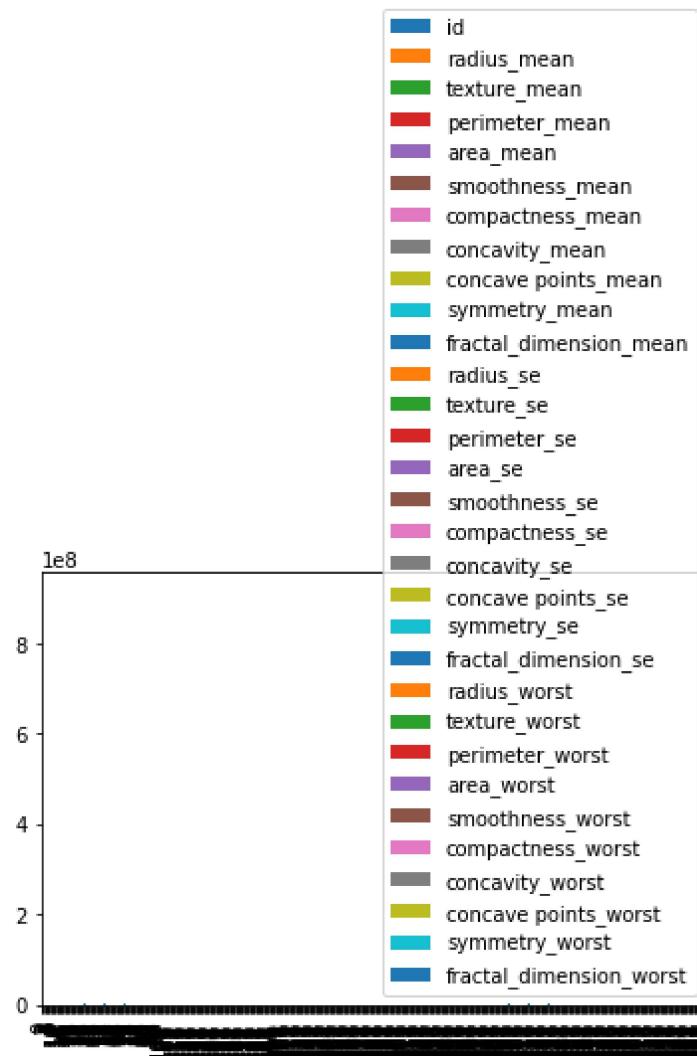
```
Out[15]: <AxesSubplot:xlabel='texture_mean', ylabel='perimeter_mean'>
```



## Bar Chart

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

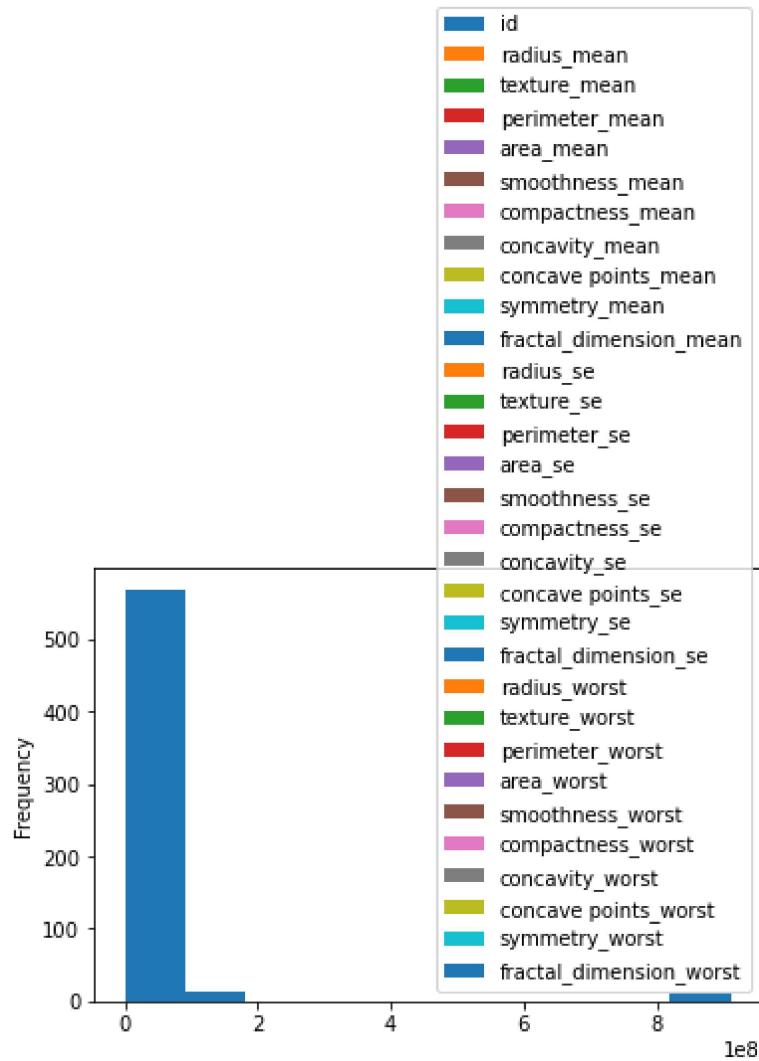
```
Out[16]: <AxesSubplot:>
```



## Histogram

```
In [17]: data.plot.hist()
```

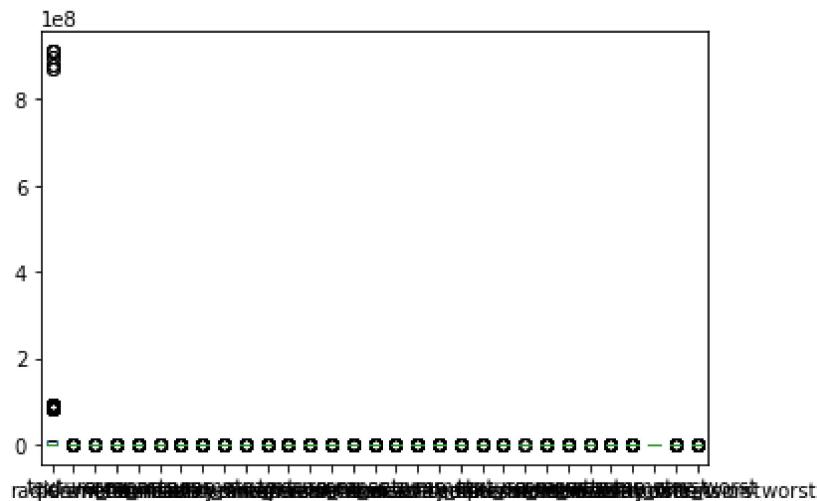
```
Out[17]: <AxesSubplot:ylabel='Frequency'>
```



## Box Plot

```
In [32]: data.plot.box('perimeter_mean')
```

```
Out[32]: <AxesSubplot:>
```

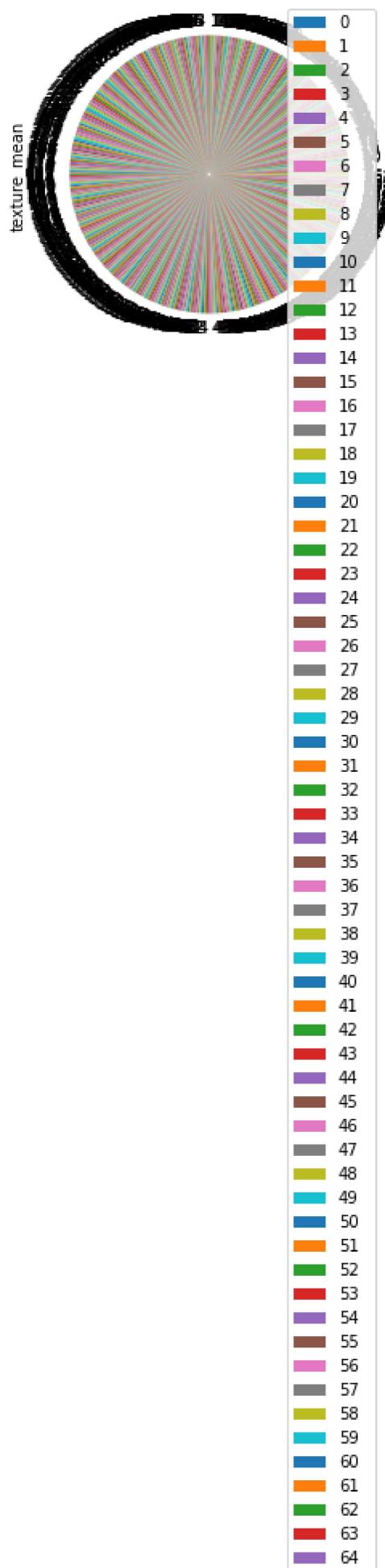


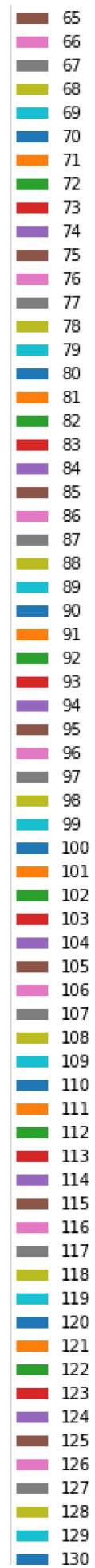
## Pie Chart

```
In [19]: data.plot.pie(y="texture_mean")
```

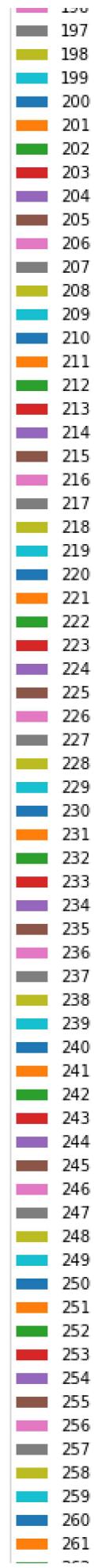
```
Out[19]: <AxesSubplot:ylabel='texture_mean'>
```







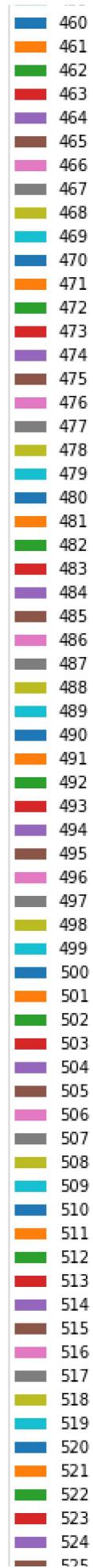
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196



262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327

328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393

394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459

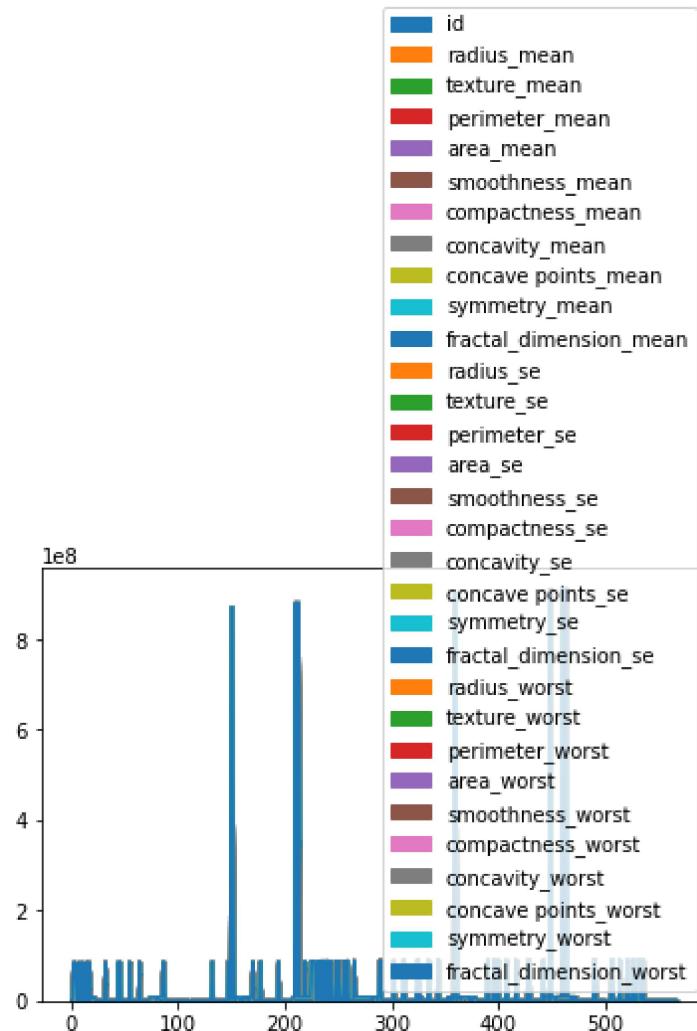




## Area

```
In [20]: data.plot.area()
```

```
Out[20]: <AxesSubplot:>
```



## To Find Mean

```
In [21]: data.mean()
```

```
Out[21]: id                  3.037183e+07
radius_mean          1.412729e+01
texture_mean          1.928965e+01
perimeter_mean        9.196903e+01
area_mean              6.548891e+02
smoothness_mean       9.636028e-02
compactness_mean      1.043410e-01
concavity_mean        8.879932e-02
concave_points_mean   4.891915e-02
symmetry_mean         1.811619e-01
fractal_dimension_mean 6.279761e-02
radius_se               4.051721e-01
texture_se              1.216853e+00
perimeter_se            2.866059e+00
area_se                 4.033708e+01
smoothness_se           7.040979e-03
compactness_se          2.547814e-02
concavity_se             3.189372e-02
concave_points_se       1.179614e-02
symmetry_se              2.054230e-02
fractal_dimension_se    3.794904e-03
radius_worst             1.626919e+01
texture_worst            2.567722e+01
perimeter_worst          1.072612e+02
area_worst                8.805831e+02
smoothness_worst         1.323686e-01
compactness_worst        2.542650e-01
concavity_worst          2.721885e-01
concave_points_worst     1.146062e-01
symmetry_worst            2.900756e-01
fractal_dimension_worst   8.394582e-02
dtype: float64
```

## To Find Median

```
In [22]: data.median()
```

```
Out[22]: id          906024.000000
radius_mean      13.370000
texture_mean     18.840000
perimeter_mean   86.240000
area_mean        551.100000
smoothness_mean  0.095870
compactness_mean 0.092630
concavity_mean   0.061540
concave_points_mean 0.033500
symmetry_mean    0.179200
fractal_dimension_mean 0.061540
radius_se         0.324200
texture_se        1.108000
perimeter_se     2.287000
area_se           24.530000
smoothness_se    0.006380
compactness_se   0.020450
concavity_se     0.025890
concave_points_se 0.010930
...
```

## To Find Mode

```
In [23]: data.mode()
```

```
Out[23]:
```

		id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	...
0		8670	B	12.34	14.93	82.61	512.2	0.095870	
1		8913	NaN	NaN	15.70	87.76	NaN	0.006380	
2		8915	NaN	NaN	16.84	134.70	NaN	0.020450	
3		9047	NaN	NaN	16.85	NaN	NaN	0.025890	
4		85715	NaN	NaN	17.46	NaN	NaN	0.010930	
...		...	...	...	...	...	...	...	...
564		911157302	NaN	NaN	NaN	NaN	NaN	NaN	
565		911296201	NaN	NaN	NaN	NaN	NaN	NaN	
566		911296202	NaN	NaN	NaN	NaN	NaN	NaN	
567		911320501	NaN	NaN	NaN	NaN	NaN	NaN	

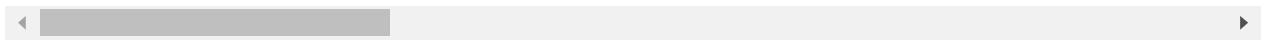
## Describe

In [24]: `data.describe()`

Out[24]:

	<code>id</code>	<code>radius_mean</code>	<code>texture_mean</code>	<code>perimeter_mean</code>	<code>area_mean</code>	<code>smoothness_mean</code>	<code>compactness_mean</code>	<code>concavity_mean</code>	<code>concave points_mean</code>	<code>symmetry_mean</code>	<code>fractal_dimension_mean</code>	<code>radius_se</code>	<code>texture_se</code>	<code>perimeter_se</code>	<code>area_se</code>	<code>smoothness_se</code>	<code>compactness_se</code>	<code>concavity_se</code>	<code>concave points_se</code>	<code>symmetry_se</code>	<code>fractal_dimension_se</code>	<code>radius_worst</code>	<code>texture_worst</code>	<code>perimeter_worst</code>	<code>area_worst</code>	<code>smoothness_worst</code>	<code>compactness_worst</code>	<code>concavity_worst</code>	<code>concave points_worst</code>	<code>symmetry_worst</code>	<code>fractal_dimension_worst</code>
<b>count</b>	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000		
<b>mean</b>	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360																									
<b>std</b>	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064																									
<b>min</b>	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630																									
<b>25%</b>	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370																									
<b>50%</b>	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870																									
<b>75%</b>	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300																									
<b>max</b>	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400																									

8 rows × 31 columns



## Sum

```
In [25]: data.sum()
```

```
Out[25]: id                                     17281572085
diagnosis                                MMMMMMMMMMMMMMMMMMBBMMMMMMMMMMBMMMMMM...
radius_mean                                 8038.429
texture_mean                                10975.81
perimeter_mean                             52330.38
area_mean                                    372631.9
smoothness_mean                            54.829
compactness_mean                           59.37002
concavity_mean                             50.526811
concave_points_mean                      27.834994
symmetry_mean                             103.0811
fractal_dimension_mean                   35.73184
radius_se                                    230.5429
texture_se                                    692.3896
perimeter_se                               1630.7877
area_se                                       22951.798
smoothness_se                             4.006317
compactness_se                            14.497061
concavity_se                               18.147525
concave_points_se                         6.712002
symmetry_se                                11.688568
fractal_dimension_se                      2.1593
radius_worst                                9257.169
texture_worst                               14610.34
perimeter_worst                            61031.63
area_worst                                    501051.8
smoothness_worst                           75.31773
compactness_worst                          144.67681
concavity_worst                            154.875247
concave_points_worst                      65.210941
symmetry_worst                             165.053
fractal_dimension_worst                   47.76517
dtype: object
```

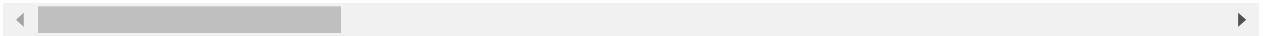
## Cumulative Sum

In [26]: `data.cumsum()`

Out[26]:

	<b>id</b>		<b>diagnosis</b>	<b>radius</b>
<b>0</b>	842302		M	
<b>1</b>	1684819		MM	
<b>2</b>	85985722		MMM	
<b>3</b>	170334023		MMMM	
<b>4</b>	254692425		MMMMM	
...	...		...	
<b>564</b>	17278698457	MMMMMM	BBBBB	7
<b>565</b>	17279625139	MMMMMM	BBBB	7
<b>566</b>	17280552093	MMMMMM	BBBB	8
<b>567</b>	17281479334	MMMMMM	BBBB	8
<b>568</b>	17281572085	MMMMMM	BBBB	8

569 rows × 32 columns



## Minimum Values

```
In [27]: data.min()
```

```
Out[27]: id                      8670
diagnosis                  B
radius_mean                 6.981
texture_mean                9.71
perimeter_mean              43.79
area_mean                   143.5
smoothness_mean             0.05263
compactness_mean            0.01938
concavity_mean               0.0
concave points_mean         0.0
symmetry_mean                0.106
fractal_dimension_mean      0.04996
radius_se                    0.1115
texture_se                   0.3602
perimeter_se                 0.757
area_se                      6.802
smoothness_se                0.001713
compactness_se               0.002252
concavity_se                 0.0
concave points_se            0.0
symmetry_se                  0.007882
fractal_dimension_se         0.000895
radius_worst                 7.93
texture_worst                 12.02
perimeter_worst              50.41
area_worst                   185.2
smoothness_worst              0.07117
compactness_worst             0.02729
concavity_worst               0.0
concave points_worst          0.0
symmetry_worst                 0.1565
fractal_dimension_worst       0.05504
dtype: object
```

## Maximum Values

```
In [28]: data.max()
```

```
Out[28]: id                      911320502
diagnosis                  M
radius_mean                 28.11
texture_mean                39.28
perimeter_mean              188.5
area_mean                   2501.0
smoothness_mean             0.1634
compactness_mean            0.3454
concavity_mean              0.4268
concave points_mean         0.2012
symmetry_mean               0.304
fractal_dimension_mean      0.09744
radius_se                    2.873
texture_se                   4.885
perimeter_se                21.98
area_se                      542.2
smoothness_se                0.03113
compactness_se               0.1354
concavity_se                 0.396
concave points_se            0.05279
symmetry_se                  0.07895
fractal_dimension_se         0.02984
radius_worst                 36.04
texture_worst                49.54
perimeter_worst              251.2
area_worst                   4254.0
smoothness_worst             0.2226
compactness_worst            1.058
concavity_worst              1.252
concave points_worst          0.291
symmetry_worst                0.6638
fractal_dimension_worst       0.2075
dtype: object
```

## Correlation

```
In [29]: from scipy.stats import spearmanr
print(spearmanr(data['texture_mean'], data['perimeter_mean']))
```

```
SpearmanResult(correlation=0.34814189073942986, pvalue=1.1756333023599274e-17)
```

## Covariance

```
In [30]: from scipy.stats import pearsonr  
print(pearsonr(data['texture_mean'], data['perimeter_mean']))  
  
(0.3295330586865702, 7.0419612377641145e-16)
```

## Count

```
In [31]: data.count()  
  
Out[31]: id                569  
diagnosis          569  
radius_mean        569  
texture_mean       569  
perimeter_mean    569  
area_mean          569  
smoothness_mean   569  
compactness_mean  569  
concavity_mean    569  
concave_points_mean 569  
symmetry_mean     569  
fractal_dimension_mean 569  
radius_se          569  
texture_se         569  
perimeter_se      569  
area_se            569  
smoothness_se     569  
compactness_se    569  
concavity_se      569  
concave_points_se 569  
symmetry_se       569  
fractal_dimension_se 569  
radius_worst       569  
texture_worst      569  
perimeter_worst   569  
area_worst         569  
smoothness_worst  569  
compactness_worst 569  
concavity_worst   569  
concave_points_worst 569  
symmetry_worst    569  
fractal_dimension_worst 569  
dtype: int64
```