SUMESH R - 20104169

Basic Analysis using NumPy and Pandas

Import Libraries

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

In [2]:
```
import numpy as np
```

In [3]:
```
from numpy import cov
from scipy.stats import pearsonr
from scipy.stats import spearmanr
```

Import Dataset

In [4]:
```
data = pd.read_csv("8_BreastCancerPrediction.csv")
```

In [5]:
```
display(data)
```

	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
...	...	...	...	...	...	...	...
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

To display top 10 rows

In [6]:
```
data.head(10)
```

Out[6]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cor
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 × 33 columns

to display last 5 rows

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

Out[7]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	cor
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 × 33 columns

statistical summary

In [8]:
```
data.describe()
```

Out[8]:

	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

To print number of elements

In [9]:
```
data.size
```

Out[9]: 18777

to print number of row and cols

In [10]:
```
data.shape
```

Out[10]: (569, 33)

to find missing values

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

Out[11]:

	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	

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	comp
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

fill null values with a constant

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

Out[12]:

	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

mean

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

Out[13]:
```
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
Unnamed: 32                         NaN
dtype: float64
```

median

In [14]:
```
data.median()
```

Out[14]:
```
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
symmetry_se                     0.018730
fractal_dimension_se            0.003187
radius_worst                   14.970000
texture_worst                  25.410000
perimeter_worst                97.660000
area_worst                    686.500000
smoothness_worst                0.131300
compactness_worst               0.211900
concavity_worst                 0.226700
concave points_worst            0.099930
symmetry_worst                  0.282200
fractal_dimension_worst         0.080040
```

```
Unnamed: 32                      NaN
dtype: float64
```

mode

In [15]:
```
data.mode()
```

Out[15]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	8670	B	12.34	14.93	82.61	512.2	0.1007
1	8913	NaN	NaN	15.70	87.76	NaN	NaN
2	8915	NaN	NaN	16.84	134.70	NaN	NaN
3	9047	NaN	NaN	16.85	NaN	NaN	NaN
4	85715	NaN	NaN	17.46	NaN	NaN	NaN
...	...	...	...	...	...	...	...
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
568	911320502	NaN	NaN	NaN	NaN	NaN	NaN

569 rows × 33 columns

cumsum

In [16]:
```
data.cumsum()
```

Out[16]:

	id	diagnosis	ra
0	842302	M	
1	1684819	MM	
2	85985722	MMM	
3	170334023	MMMM	
4	254692425	MMMMM	
...	...	...	
564	17278698457	MMMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMMBMMMMMMMM...	
565	17279625139	MMMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMMBMMMMMMMM...	
566	17280552093	MMMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMMBMMMMMMMM...	

	id	diagnosis	ra
567	17281479334	MMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMBMMMMMMMM...	
568	17281572085	MMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMBMMMMMMMM...	

569 rows × 33 columns

min

In [17]:
```
data.min()
```

Out[17]:
```
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
Unnamed: 32                   NaN
dtype: object
```

max

In [18]:
```
data.max()
```

Out[18]:
```
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
Unnamed: 32                    NaN
dtype: object
```

sum

In [19]:
```
data.sum()
```

Out[19]:
```
id                                                         17281572085
diagnosis                 MMMMMMMMMMMMMMMMMMMBBBMMMMMMMMMMMMMMMMMBMMMMMMMM...
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
Unnamed: 32                                                           0.0
dtype: object
```

count

In [20]:
```
data.count()
```

Out[20]:
```
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
Unnamed: 32                  0
dtype: int64
```

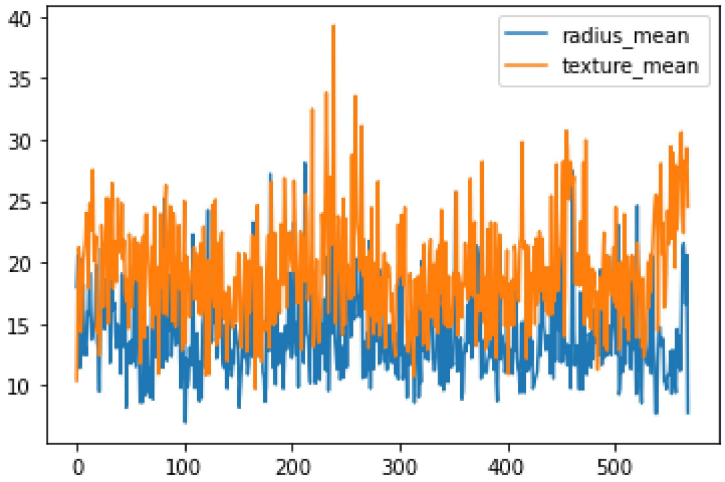
to select a particular columns

In [21]:
```
df=pd.DataFrame(data[['radius_mean','texture_mean']])
import matplotlib.pyplot as plt
```

line plot

In [22]:
```
df.plot.line()
```

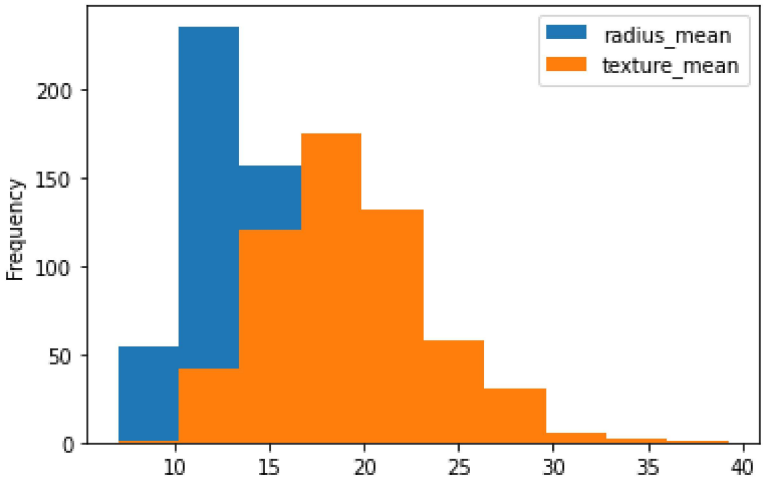
Out[22]:
```
<AxesSubplot:>
```



histogram

In [23]:
```
df.plot.hist()
```

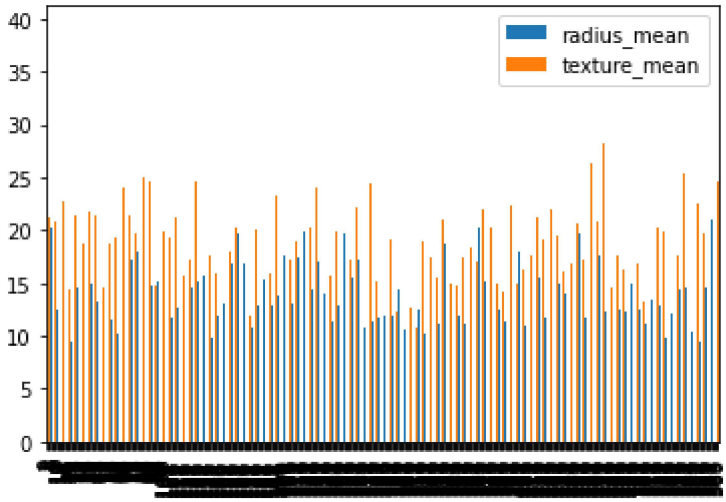
Out[23]: <AxesSubplot:ylabel='Frequency'>



bar chart

In [24]:
```
df.plot.bar()
```

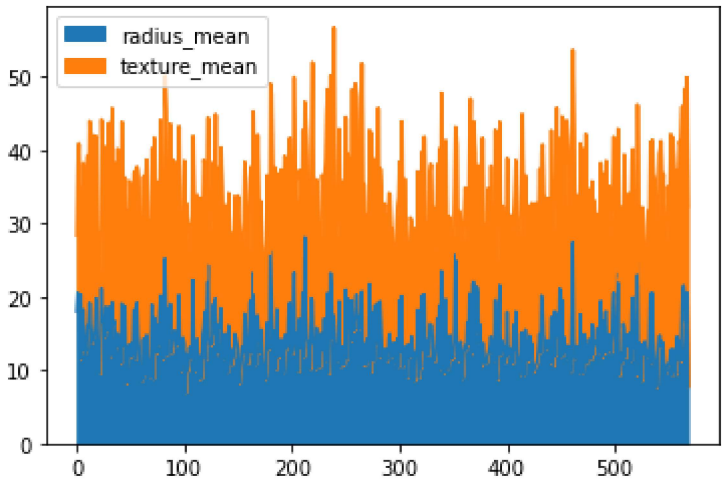
Out[24]: <AxesSubplot:>



area plot

In [25]:
```
df.plot.area()
```

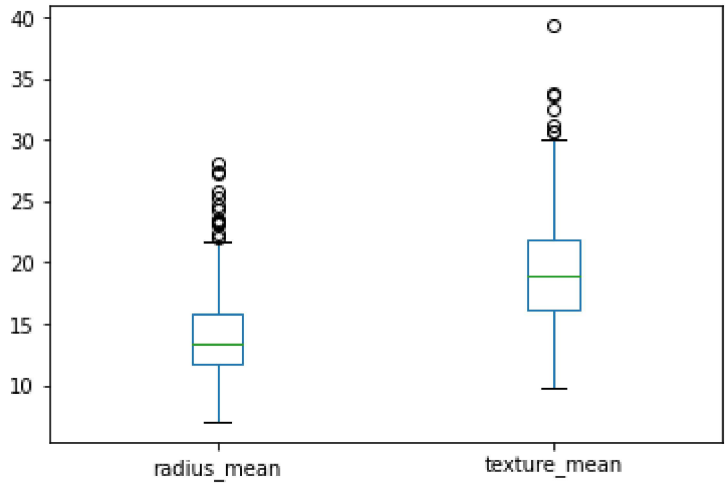
Out[25]: <AxesSubplot:>



box plot

In [26]:
```
df.plot.box()
```

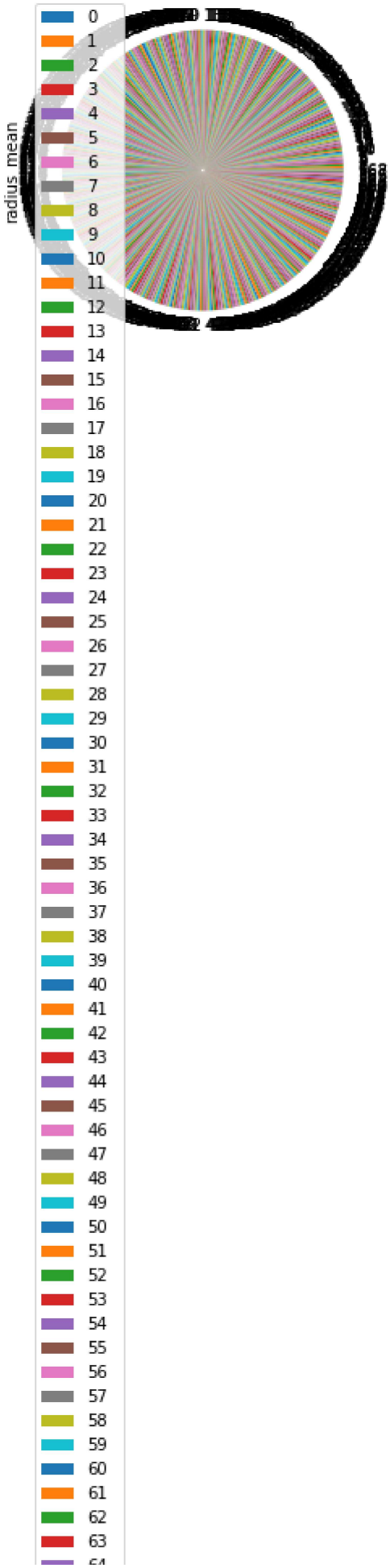
Out[26]: <AxesSubplot:>

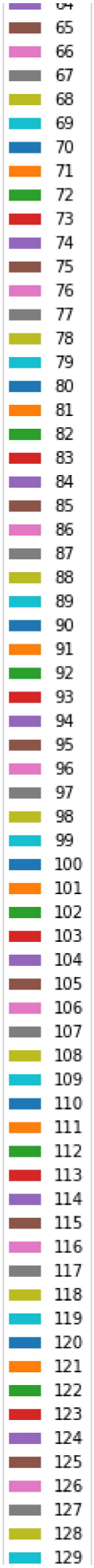


pie plot

In [27]:
```
df.plot.pie(y="radius_mean")
```

Out[27]:
```
<AxesSubplot:ylabel='radius_mean'>
```

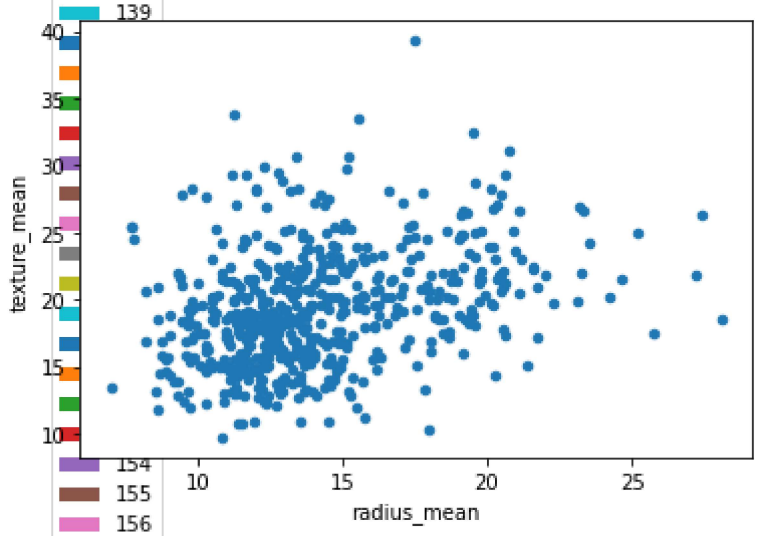

radius_mean
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63


64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129

scatter plot

In [28]:
```
df.plot.scatter(x="radius_mean",y="texture_mean")
```

Out[28]:
```
<AxesSubplot:xlabel='radius_mean', ylabel='texture_mean'>
```



covariance

In [29]:
```
cov(data["radius_mean"],data["texture_mean"])
```

Out[29]:
```
array([[12.41892013,  4.90758156],
       [ 4.90758156, 18.49890868]])
```

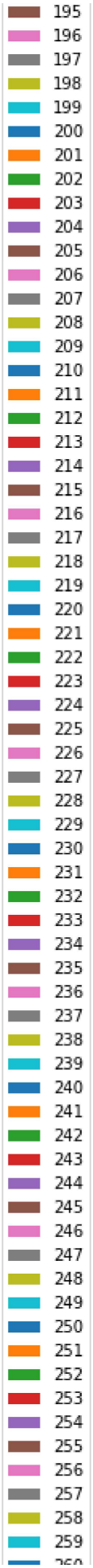
correlation

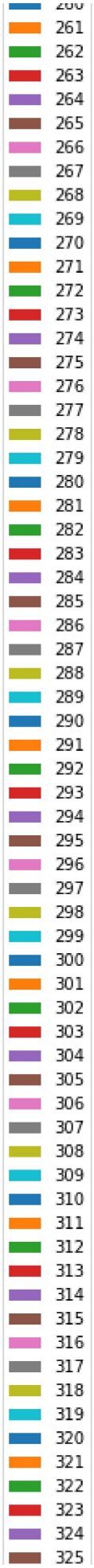
In [30]:
```
spearmanr(data["radius_mean"],data["texture_mean"])
```

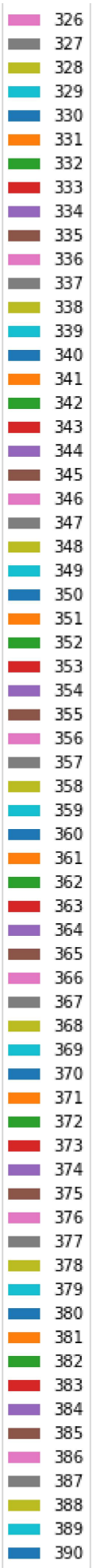
Out[30]:
```
SpearmanrResult(correlation=0.3409562685372812, pvalue=5.900189597213798e-17)
```

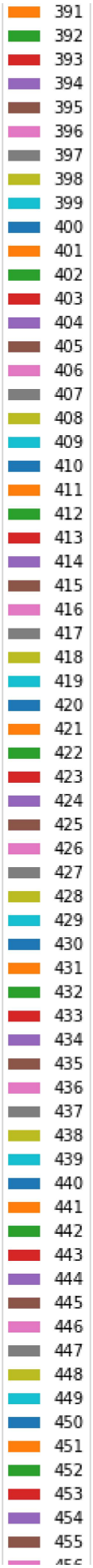
In [31]:
```
pearsonr(data["radius_mean"],data["texture_mean"])
```

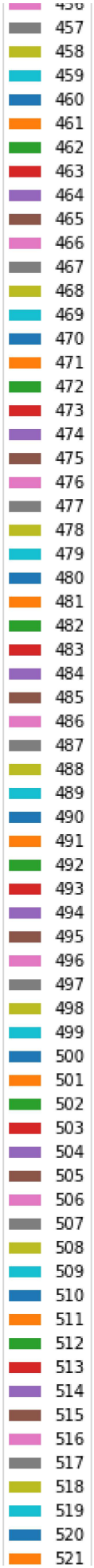
Out[31]:
```
(0.3237810890927733, 2.360374375922593e-15)
```

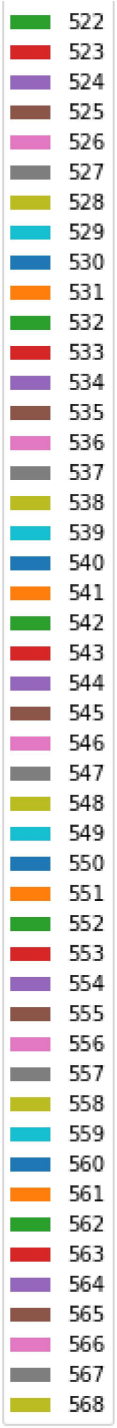

195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259


260
261
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
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521


522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568