K_Means

March 19, 2024

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

[ ]:
```
df = pd.read_csv("datasets/breast-cancer.csv")
df.head()
```

[ ]:

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

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.3001	0.14710	
1	0.08474	0.07864	0.0869	0.07017	
2	0.10960	0.15990	0.1974	0.12790	
3	0.14250	0.28390	0.2414	0.10520	
4	0.10030	0.13280	0.1980	0.10430	

	…	radius_worst	texture_worst	perimeter_worst	area_worst	\
0	…	25.38	17.33	184.60	2019.0	
1	…	24.99	23.41	158.80	1956.0	
2	…	23.57	25.53	152.50	1709.0	
3	…	14.91	26.50	98.87	567.7	
4	…	22.54	16.67	152.20	1575.0	

	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	\
0	0.1622	0.6656	0.7119	0.2654	
1	0.1238	0.1866	0.2416	0.1860	
2	0.1444	0.4245	0.4504	0.2430	
3	0.2098	0.8663	0.6869	0.2575	
4	0.1374	0.2050	0.4000	0.1625	

```
   symmetry_worst  fractal_dimension_worst
0          0.4601                  0.11890
1          0.2750                  0.08902
2          0.3613                  0.08758
3          0.6638                  0.17300
4          0.2364                  0.07678

[5 rows x 32 columns]
```

1 Drop Unnecessary columns & Basic Exploration

[ ]:
```
df = df.drop(columns=["id"])
df.head()
```

[ ]:
```
  diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0         M        17.99         10.38          122.80     1001.0
1         M        20.57         17.77          132.90     1326.0
2         M        19.69         21.25          130.00     1203.0
3         M        11.42         20.38           77.58      386.1
4         M        20.29         14.34          135.10     1297.0

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
0         0.2419  …         25.38          17.33           184.60
1         0.1812  …         24.99          23.41           158.80
2         0.2069  …         23.57          25.53           152.50
3         0.2597  …         14.91          26.50            98.87
4         0.1809  …         22.54          16.67           152.20

   area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0      2019.0            0.1622             0.6656           0.7119
1      1956.0            0.1238             0.1866           0.2416
2      1709.0            0.1444             0.4245           0.4504
3       567.7            0.2098             0.8663           0.6869
4      1575.0            0.1374             0.2050           0.4000

   concave points_worst  symmetry_worst  fractal_dimension_worst
0                0.2654          0.4601                  0.11890
1                0.1860          0.2750                  0.08902
2                0.2430          0.3613                  0.08758
```

```
3                0.2575          0.6638                0.17300
4                0.1625          0.2364                0.07678

[5 rows x 31 columns]
```

[ ]:
```
df.shape
```

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

[ ]:
```
df.diagnosis.value_counts()
```

[ ]:
```
diagnosis
B    357
M    212
Name: count, dtype: int64
```

[ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   diagnosis                569 non-null    object
 1   radius_mean              569 non-null    float64
 2   texture_mean             569 non-null    float64
 3   perimeter_mean           569 non-null    float64
 4   area_mean                569 non-null    float64
 5   smoothness_mean          569 non-null    float64
 6   compactness_mean         569 non-null    float64
 7   concavity_mean           569 non-null    float64
 8   concave points_mean      569 non-null    float64
 9   symmetry_mean            569 non-null    float64
 10  fractal_dimension_mean   569 non-null    float64
 11  radius_se                569 non-null    float64
 12  texture_se               569 non-null    float64
 13  perimeter_se             569 non-null    float64
 14  area_se                  569 non-null    float64
 15  smoothness_se            569 non-null    float64
 16  compactness_se           569 non-null    float64
 17  concavity_se             569 non-null    float64
 18  concave points_se        569 non-null    float64
 19  symmetry_se              569 non-null    float64
 20  fractal_dimension_se     569 non-null    float64
 21  radius_worst             569 non-null    float64
 22  texture_worst            569 non-null    float64
 23  perimeter_worst          569 non-null    float64
 24  area_worst               569 non-null    float64
```

```
 25  smoothness_worst         569 non-null    float64
 26  compactness_worst        569 non-null    float64
 27  concavity_worst          569 non-null    float64
 28  concave points_worst     569 non-null    float64
 29  symmetry_worst           569 non-null    float64
 30  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), object(1)
memory usage: 137.9+ KB
```

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

[ ]:

	radius_mean	texture_mean	perimeter_mean	area_mean	\
count	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	
std	3.524049	4.301036	24.298981	351.914129	
min	6.981000	9.710000	43.790000	143.500000	
25%	11.700000	16.170000	75.170000	420.300000	
50%	13.370000	18.840000	86.240000	551.100000	
75%	15.780000	21.800000	104.100000	782.700000	
max	28.110000	39.280000	188.500000	2501.000000	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
count	569.000000	569.000000	569.000000	569.000000	
mean	0.096360	0.104341	0.088799	0.048919	
std	0.014064	0.052813	0.079720	0.038803	
min	0.052630	0.019380	0.000000	0.000000	
25%	0.086370	0.064920	0.029560	0.020310	
50%	0.095870	0.092630	0.061540	0.033500	
75%	0.105300	0.130400	0.130700	0.074000	
max	0.163400	0.345400	0.426800	0.201200	

	symmetry_mean	fractal_dimension_mean	…	radius_worst	\
count	569.000000	569.000000	…	569.000000	
mean	0.181162	0.062798	…	16.269190	
std	0.027414	0.007060	…	4.833242	
min	0.106000	0.049960	…	7.930000	
25%	0.161900	0.057700	…	13.010000	
50%	0.179200	0.061540	…	14.970000	
75%	0.195700	0.066120	…	18.790000	
max	0.304000	0.097440	…	36.040000	

	texture_worst	perimeter_worst	area_worst	smoothness_worst	\
count	569.000000	569.000000	569.000000	569.000000	
mean	25.677223	107.261213	880.583128	0.132369	
std	6.146258	33.602542	569.356993	0.022832	
min	12.020000	50.410000	185.200000	0.071170	
25%	21.080000	84.110000	515.300000	0.116600	

```
50%        25.410000        97.660000   686.500000          0.131300
75%        29.720000       125.400000  1084.000000          0.146000
max        49.540000       251.200000  4254.000000          0.222600

       compactness_worst  concavity_worst  concave points_worst  \
count         569.000000       569.000000            569.000000
mean            0.254265         0.272188              0.114606
std             0.157336         0.208624              0.065732
min             0.027290         0.000000              0.000000
25%             0.147200         0.114500              0.064930
50%             0.211900         0.226700              0.099930
75%             0.339100         0.382900              0.161400
max             1.058000         1.252000              0.291000

       symmetry_worst  fractal_dimension_worst
count      569.000000               569.000000
mean         0.290076                 0.083946
std          0.061867                 0.018061
min          0.156500                 0.055040
25%          0.250400                 0.071460
50%          0.282200                 0.080040
75%          0.317900                 0.092080
max          0.663800                 0.207500

[8 rows x 30 columns]
```

[ ]:
```
df.corr(numeric_only=True)
```

[ ]:
```
                        radius_mean  texture_mean  perimeter_mean  area_mean  \
radius_mean                1.000000      0.323782        0.997855   0.987357
texture_mean               0.323782      1.000000        0.329533   0.321086
perimeter_mean             0.997855      0.329533        1.000000   0.986507
area_mean                  0.987357      0.321086        0.986507   1.000000
smoothness_mean            0.170581     -0.023389        0.207278   0.177028
compactness_mean           0.506124      0.236702        0.556936   0.498502
concavity_mean             0.676764      0.302418        0.716136   0.685983
concave points_mean        0.822529      0.293464        0.850977   0.823269
symmetry_mean              0.147741      0.071401        0.183027   0.151293
fractal_dimension_mean    -0.311631     -0.076437       -0.261477  -0.283110
radius_se                  0.679090      0.275869        0.691765   0.732562
texture_se                -0.097317      0.386358       -0.086761  -0.066280
perimeter_se               0.674172      0.281673        0.693135   0.726628
area_se                    0.735864      0.259845        0.744983   0.800086
smoothness_se             -0.222600      0.006614       -0.202694  -0.166777
compactness_se             0.206000      0.191975        0.250744   0.212583
concavity_se               0.194204      0.143293        0.228082   0.207660
concave points_se          0.376169      0.163851        0.407217   0.372320
```

symmetry_se	-0.104321	0.009127	-0.081629	-0.072497
fractal_dimension_se	-0.042641	0.054458	-0.005523	-0.019887
radius_worst	0.969539	0.352573	0.969476	0.962746
texture_worst	0.297008	0.912045	0.303038	0.287489
perimeter_worst	0.965137	0.358040	0.970387	0.959120
area_worst	0.941082	0.343546	0.941550	0.959213
smoothness_worst	0.119616	0.077503	0.150549	0.123523
compactness_worst	0.413463	0.277830	0.455774	0.390410
concavity_worst	0.526911	0.301025	0.563879	0.512606
concave points_worst	0.744214	0.295316	0.771241	0.722017
symmetry_worst	0.163953	0.105008	0.189115	0.143570
fractal_dimension_worst	0.007066	0.119205	0.051019	0.003738

	smoothness_mean	compactness_mean	concavity_mean	\
radius_mean	0.170581	0.506124	0.676764	
texture_mean	-0.023389	0.236702	0.302418	
perimeter_mean	0.207278	0.556936	0.716136	
area_mean	0.177028	0.498502	0.685983	
smoothness_mean	1.000000	0.659123	0.521984	
compactness_mean	0.659123	1.000000	0.883121	
concavity_mean	0.521984	0.883121	1.000000	
concave points_mean	0.553695	0.831135	0.921391	
symmetry_mean	0.557775	0.602641	0.500667	
fractal_dimension_mean	0.584792	0.565369	0.336783	
radius_se	0.301467	0.497473	0.631925	
texture_se	0.068406	0.046205	0.076218	
perimeter_se	0.296092	0.548905	0.660391	
area_se	0.246552	0.455653	0.617427	
smoothness_se	0.332375	0.135299	0.098564	
compactness_se	0.318943	0.738722	0.670279	
concavity_se	0.248396	0.570517	0.691270	
concave points_se	0.380676	0.642262	0.683260	
symmetry_se	0.200774	0.229977	0.178009	
fractal_dimension_se	0.283607	0.507318	0.449301	
radius_worst	0.213120	0.535315	0.688236	
texture_worst	0.036072	0.248133	0.299879	
perimeter_worst	0.238853	0.590210	0.729565	
area_worst	0.206718	0.509604	0.675987	
smoothness_worst	0.805324	0.565541	0.448822	
compactness_worst	0.472468	0.865809	0.754968	
concavity_worst	0.434926	0.816275	0.884103	
concave points_worst	0.503053	0.815573	0.861323	
symmetry_worst	0.394309	0.510223	0.409464	
fractal_dimension_worst	0.499316	0.687382	0.514930	

	concave points_mean	symmetry_mean	\
radius_mean	0.822529	0.147741	

```
texture_mean                      0.293464       0.071401
perimeter_mean                    0.850977       0.183027
area_mean                         0.823269       0.151293
smoothness_mean                   0.553695       0.557775
compactness_mean                  0.831135       0.602641
concavity_mean                    0.921391       0.500667
concave points_mean               1.000000       0.462497
symmetry_mean                     0.462497       1.000000
fractal_dimension_mean            0.166917       0.479921
radius_se                         0.698050       0.303379
texture_se                        0.021480       0.128053
perimeter_se                      0.710650       0.313893
area_se                           0.690299       0.223970
smoothness_se                     0.027653       0.187321
compactness_se                    0.490424       0.421659
concavity_se                      0.439167       0.342627
concave points_se                 0.615634       0.393298
symmetry_se                       0.095351       0.449137
fractal_dimension_se              0.257584       0.331786
radius_worst                      0.830318       0.185728
texture_worst                     0.292752       0.090651
perimeter_worst                   0.855923       0.219169
area_worst                        0.809630       0.177193
smoothness_worst                  0.452753       0.426675
compactness_worst                 0.667454       0.473200
concavity_worst                   0.752399       0.433721
concave points_worst              0.910155       0.430297
symmetry_worst                    0.375744       0.699826
fractal_dimension_worst           0.368661       0.438413

                        fractal_dimension_mean  …  radius_worst  \
radius_mean                          -0.311631  …      0.969539
texture_mean                         -0.076437  …      0.352573
perimeter_mean                       -0.261477  …      0.969476
area_mean                            -0.283110  …      0.962746
smoothness_mean                       0.584792  …      0.213120
compactness_mean                      0.565369  …      0.535315
concavity_mean                        0.336783  …      0.688236
concave points_mean                   0.166917  …      0.830318
symmetry_mean                         0.479921  …      0.185728
fractal_dimension_mean                1.000000  …     -0.253691
radius_se                             0.000111  …      0.715065
texture_se                            0.164174  …     -0.111690
perimeter_se                          0.039830  …      0.697201
area_se                              -0.090170  …      0.757373
smoothness_se                         0.401964  …     -0.230691
compactness_se                        0.559837  …      0.204607
```

```
concavity_se                        0.446630  …      0.186904
concave points_se                   0.341198  …      0.358127
symmetry_se                         0.345007  …     -0.128121
fractal_dimension_se                0.688132  …     -0.037488
radius_worst                       -0.253691  …      1.000000
texture_worst                      -0.051269  …      0.359921
perimeter_worst                    -0.205151  …      0.993708
area_worst                         -0.231854  …      0.984015
smoothness_worst                    0.504942  …      0.216574
compactness_worst                   0.458798  …      0.475820
concavity_worst                     0.346234  …      0.573975
concave points_worst                0.175325  …      0.787424
symmetry_worst                      0.334019  …      0.243529
fractal_dimension_worst             0.767297  …      0.093492

                        texture_worst  perimeter_worst  area_worst  \
radius_mean                  0.297008         0.965137    0.941082
texture_mean                 0.912045         0.358040    0.343546
perimeter_mean               0.303038         0.970387    0.941550
area_mean                    0.287489         0.959120    0.959213
smoothness_mean              0.036072         0.238853    0.206718
compactness_mean             0.248133         0.590210    0.509604
concavity_mean               0.299879         0.729565    0.675987
concave points_mean          0.292752         0.855923    0.809630
symmetry_mean                0.090651         0.219169    0.177193
fractal_dimension_mean      -0.051269        -0.205151   -0.231854
radius_se                    0.194799         0.719684    0.751548
texture_se                   0.409003        -0.102242   -0.083195
perimeter_se                 0.200371         0.721031    0.730713
area_se                      0.196497         0.761213    0.811408
smoothness_se               -0.074743        -0.217304   -0.182195
compactness_se               0.143003         0.260516    0.199371
concavity_se                 0.100241         0.226680    0.188353
concave points_se            0.086741         0.394999    0.342271
symmetry_se                 -0.077473        -0.103753   -0.110343
fractal_dimension_se        -0.003195        -0.001000   -0.022736
radius_worst                 0.359921         0.993708    0.984015
texture_worst                1.000000         0.365098    0.345842
perimeter_worst              0.365098         1.000000    0.977578
area_worst                   0.345842         0.977578    1.000000
smoothness_worst             0.225429         0.236775    0.209145
compactness_worst            0.360832         0.529408    0.438296
concavity_worst              0.368366         0.618344    0.543331
concave points_worst         0.359755         0.816322    0.747419
symmetry_worst               0.233027         0.269493    0.209146
fractal_dimension_worst      0.219122         0.138957    0.079647
```

	smoothness_worst	compactness_worst	concavity_worst	\
radius_mean	0.119616	0.413463	0.526911	
texture_mean	0.077503	0.277830	0.301025	
perimeter_mean	0.150549	0.455774	0.563879	
area_mean	0.123523	0.390410	0.512606	
smoothness_mean	0.805324	0.472468	0.434926	
compactness_mean	0.565541	0.865809	0.816275	
concavity_mean	0.448822	0.754968	0.884103	
concave points_mean	0.452753	0.667454	0.752399	
symmetry_mean	0.426675	0.473200	0.433721	
fractal_dimension_mean	0.504942	0.458798	0.346234	
radius_se	0.141919	0.287103	0.380585	
texture_se	-0.073658	-0.092439	-0.068956	
perimeter_se	0.130054	0.341919	0.418899	
area_se	0.125389	0.283257	0.385100	
smoothness_se	0.314457	-0.055558	-0.058298	
compactness_se	0.227394	0.678780	0.639147	
concavity_se	0.168481	0.484858	0.662564	
concave points_se	0.215351	0.452888	0.549592	
symmetry_se	-0.012662	0.060255	0.037119	
fractal_dimension_se	0.170568	0.390159	0.379975	
radius_worst	0.216574	0.475820	0.573975	
texture_worst	0.225429	0.360832	0.368366	
perimeter_worst	0.236775	0.529408	0.618344	
area_worst	0.209145	0.438296	0.543331	
smoothness_worst	1.000000	0.568187	0.518523	
compactness_worst	0.568187	1.000000	0.892261	
concavity_worst	0.518523	0.892261	1.000000	
concave points_worst	0.547691	0.801080	0.855434	
symmetry_worst	0.493838	0.614441	0.532520	
fractal_dimension_worst	0.617624	0.810455	0.686511	

	concave points_worst	symmetry_worst	\
radius_mean	0.744214	0.163953	
texture_mean	0.295316	0.105008	
perimeter_mean	0.771241	0.189115	
area_mean	0.722017	0.143570	
smoothness_mean	0.503053	0.394309	
compactness_mean	0.815573	0.510223	
concavity_mean	0.861323	0.409464	
concave points_mean	0.910155	0.375744	
symmetry_mean	0.430297	0.699826	
fractal_dimension_mean	0.175325	0.334019	
radius_se	0.531062	0.094543	
texture_se	-0.119638	-0.128215	
perimeter_se	0.554897	0.109930	
area_se	0.538166	0.074126	

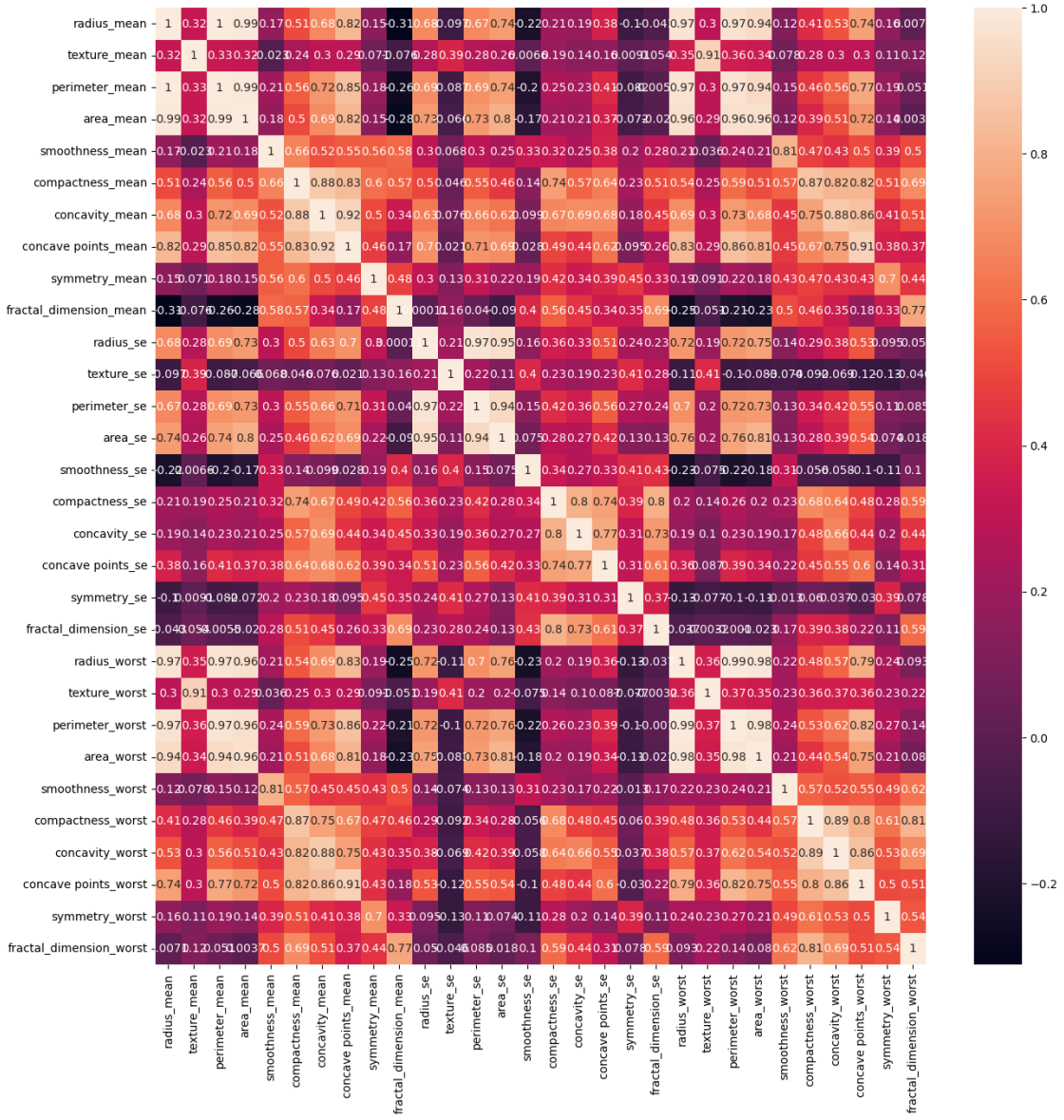
smoothness_se	-0.102007	-0.107342
compactness_se	0.483208	0.277878
concavity_se	0.440472	0.197788
concave points_se	0.602450	0.143116
symmetry_se	-0.030413	0.389402
fractal_dimension_se	0.215204	0.111094
radius_worst	0.787424	0.243529
texture_worst	0.359755	0.233027
perimeter_worst	0.816322	0.269493
area_worst	0.747419	0.209146
smoothness_worst	0.547691	0.493838
compactness_worst	0.801080	0.614441
concavity_worst	0.855434	0.532520
concave points_worst	1.000000	0.502528
symmetry_worst	0.502528	1.000000
fractal_dimension_worst	0.511114	0.537848

	fractal_dimension_worst
radius_mean	0.007066
texture_mean	0.119205
perimeter_mean	0.051019
area_mean	0.003738
smoothness_mean	0.499316
compactness_mean	0.687382
concavity_mean	0.514930
concave points_mean	0.368661
symmetry_mean	0.438413
fractal_dimension_mean	0.767297
radius_se	0.049559
texture_se	-0.045655
perimeter_se	0.085433
area_se	0.017539
smoothness_se	0.101480
compactness_se	0.590973
concavity_se	0.439329
concave points_se	0.310655
symmetry_se	0.078079
fractal_dimension_se	0.591328
radius_worst	0.093492
texture_worst	0.219122
perimeter_worst	0.138957
area_worst	0.079647
smoothness_worst	0.617624
compactness_worst	0.810455
concavity_worst	0.686511
concave points_worst	0.511114
symmetry_worst	0.537848

```
fractal_dimension_worst                 1.000000

[30 rows x 30 columns]
```

[ ]:
```
plt.figure(figsize=(15,15))
sns.heatmap(df.corr(numeric_only= True),annot=True )
```

[ ]: <Axes: >



2 Categorise Columns

[ ]:
```
num_cols = list(df.select_dtypes(include=['int', 'float']).columns)
cat_cols = list(df.select_dtypes(exclude=['int', 'float']).columns)
```

[ ]:
```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OrdinalEncoder
from sklearn.impute import SimpleImputer
from sklearn.base import BaseEstimator, TransformerMixin
```

3 Custom IQR Removal Transformer

[ ]:
```
class Outlier_Remover(BaseEstimator, TransformerMixin):

    def __init__(self, action='keep'):
        self.action = action

    def fit(self, X, y=None):
        self.median_ = np.median(X, axis=0)
        return self

    def transform(self, X):
        Q1 = np.percentile(X, 25, axis=0)
        Q3 = np.percentile(X, 75, axis=0)

        IQR = Q3 - Q1

        lower = Q1 - 1.5*IQR
        upper = Q3 + 1.5*IQR

        outlier_mask = (X < lower) | (X > upper)

        if self.action == 'drop':
            return X[~outlier_mask]
        else:
            for i in range(X.shape[1]):
                X[:, i][outlier_mask[:, i]] = self.median_[i]
            return X
```

4 Define Pipeline's for both type of columns

[ ]:
```
cat_preprocessor = Pipeline(steps=[
    ("cat_null_handler",SimpleImputer(missing_values=np.
 ↪nan,strategy="most_frequent")),
    ("cat_enocder",OrdinalEncoder()),
])
```

```
num_preprocessor = Pipeline(steps=[
    ("num_null_handler",SimpleImputer(missing_values=np.nan,strategy='median')),
    ("num_outlier_remover",Outlier_Remover(action="keep")),
    ("num_scaler",StandardScaler()),
])
```

5 Preprocess data with Pipeline

[ ]:
```
df[num_cols] = num_preprocessor.fit_transform(df[num_cols])
df[cat_cols] = cat_preprocessor.fit_transform(df[cat_cols])
df.head()
```

[ ]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	\
0	1.0	1.335706	-2.183545	1.526900	1.477828	
1	1.0	2.168713	-0.336098	1.999151	2.692856	
2	1.0	1.884587	0.533878	1.863554	2.233014	
3	1.0	-0.785558	0.316384	-0.587475	-0.821005	
4	1.0	2.078309	-1.193573	2.102017	2.584438	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	1.708051	-0.152054	-0.283578	2.887302	
1	-0.858619	-0.466377	0.107062	0.680991	
2	1.037027	1.359343	1.809179	2.336656	
3	-0.009925	-0.152054	2.486945	1.685632	
4	0.327875	0.750470	1.818421	1.659821	

	symmetry_mean	…	radius_worst	texture_worst	perimeter_worst	\
0	2.646196	…	2.313300	-1.400167	2.761488	
1	0.084394	…	2.219021	-0.357761	1.873778	
2	1.169045	…	1.875747	0.005710	1.657012	
3	-0.000015	…	-0.217742	0.172015	-0.188255	
4	0.071733	…	1.626752	-1.513323	1.646690	

	area_worst	smoothness_worst	compactness_worst	concavity_worst	\
0	-0.224151	1.429505	-0.206236	2.497158	
1	-0.224151	-0.369836	-0.404542	-0.084360	
2	2.425960	0.595436	1.460157	1.061762	
3	-0.532056	-0.018402	-0.206236	2.359931	
4	2.078659	0.267431	-0.260319	0.785112	

	concave points_worst	symmetry_worst	fractal_dimension_worst
0	2.296076	0.002570	2.739080
1	1.087084	-0.152067	0.548993
2	1.955000	1.701432	0.443446

```
3              2.175786         0.002570                -0.109206
4              0.729259        -0.981094                -0.348151

[5 rows x 31 columns]
```

6 Train Test Split

[ ]:
```
X = df.drop(columns=["diagnosis"])
y = df.diagnosis
```

[ ]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,␣
 ↪random_state=42)
```

[ ]:
```
X_train.head()
```

[ ]:
```
     radius_mean  texture_mean  perimeter_mean  area_mean  smoothness_mean  \
68     -1.557543     -0.446094       -1.466049  -1.327953         0.808268
181     2.336606      1.863840        2.457374   2.636777         1.380164
63     -1.511050     -1.313570       -1.446878  -1.289072        -1.432803
248    -1.034169      1.526349       -1.034944  -0.967183         0.043452
60     -1.189147     -1.058577       -1.196725  -1.098406         1.326787

     compactness_mean  concavity_mean  concave points_mean  symmetry_mean  \
68           0.941445       -0.283578            -0.076721       1.346303
181         -0.152054        2.599393             2.959001       2.544905
63          -0.267089       -0.309148            -0.706235       2.317002
248         -0.607923       -0.865070            -0.868274       0.443130
60          -0.422115       -1.064549            -0.961483      -0.000015

     fractal_dimension_mean  …  radius_worst  texture_worst  \
68                -0.104601  …     -1.329757      -0.488062
181                2.026677  …      2.627565       1.368725
63                 1.281415  …     -1.402279      -1.074415
248                0.195217  …     -0.860776       1.661902
60                 1.276275  …     -1.158119      -1.379593

     perimeter_worst  area_worst  smoothness_worst  compactness_worst  \
68         -1.336429   -1.161862          0.773495           1.554215
181         2.482789   -0.224151          0.815668          -0.206236
63         -1.333332   -1.199702         -1.561900          -0.551899
248        -0.907025   -0.822337          0.853154          -0.771367
60         -1.186412   -1.048082         -0.196462          -1.093829

     concavity_worst  concave points_worst  symmetry_worst  \
68         -0.166147              0.919592        0.002570
```

```
181         2.311078              2.675218        2.743086
63         -0.643698             -0.970486        0.990531
248        -0.793001             -0.810759        1.263294
60         -1.291520             -1.352369        1.581159

     fractal_dimension_worst
68                  2.636465
181                -0.109206
63                  0.247013
248                -0.004393
60                 -0.097479

[5 rows x 30 columns]
```

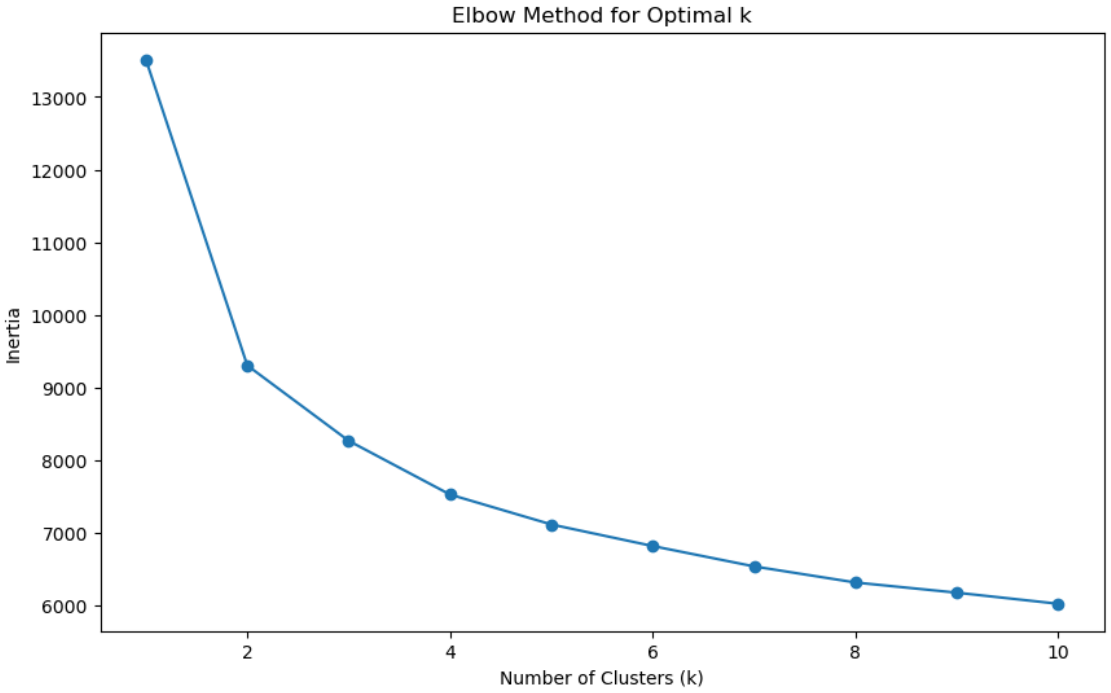
[ ]:
```
y_train.head()
```

[ ]:
```
68     0.0
181    1.0
63     0.0
248    0.0
60     0.0
Name: diagnosis, dtype: float64
```

[ ]:
```
from sklearn.cluster import KMeans
```

7 Elbow Method

[ ]:
```
inertia = []
k_range = range(1, 11)  # Test k values from 1 to 10

for k in k_range:
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X_train)
    inertia.append(kmeans.inertia_)

# Plot the Elbow Method
plt.figure(figsize=(10, 6))
plt.plot(k_range, inertia, marker='o')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal k')
plt.show()
```



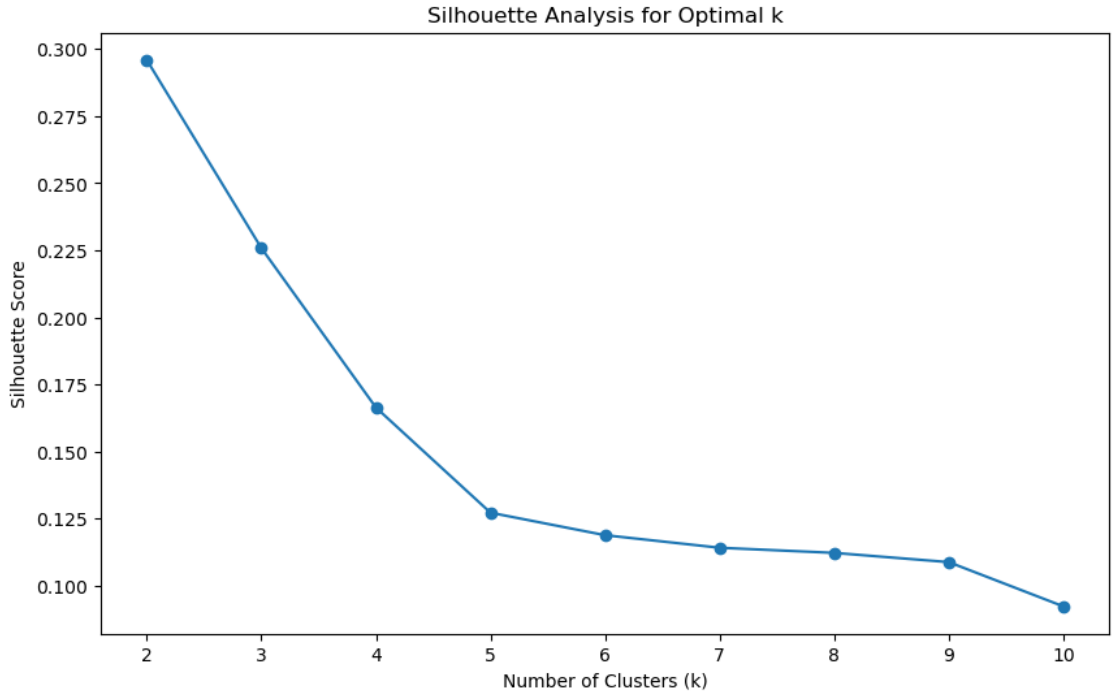
8 Silhouette Method

[ ]:
```
from sklearn.metrics import silhouette_score
```

[ ]:
```
silhouette_scores = []
k_range = range(2, 11)  # Test k values from 2 to 10

for k in k_range:
    kmeans_sil = KMeans(n_clusters=k)
    kmeans_sil.fit(X_train)
    score = silhouette_score(X_train, kmeans_sil.labels_)
    silhouette_scores.append(score)

optimal_k = k_range[silhouette_scores.index(max(silhouette_scores))]
print(f'Optimal number of clusters (k) according to silhouette analysis:␣
 ↪{optimal_k}')

# Plot the Silhouette Method
plt.figure(figsize=(10, 6))
plt.plot(k_range, silhouette_scores, marker='o')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Analysis for Optimal k')
```

```
plt.show()
```

```
Optimal number of clusters (k) according to silhouette analysis: 2
```

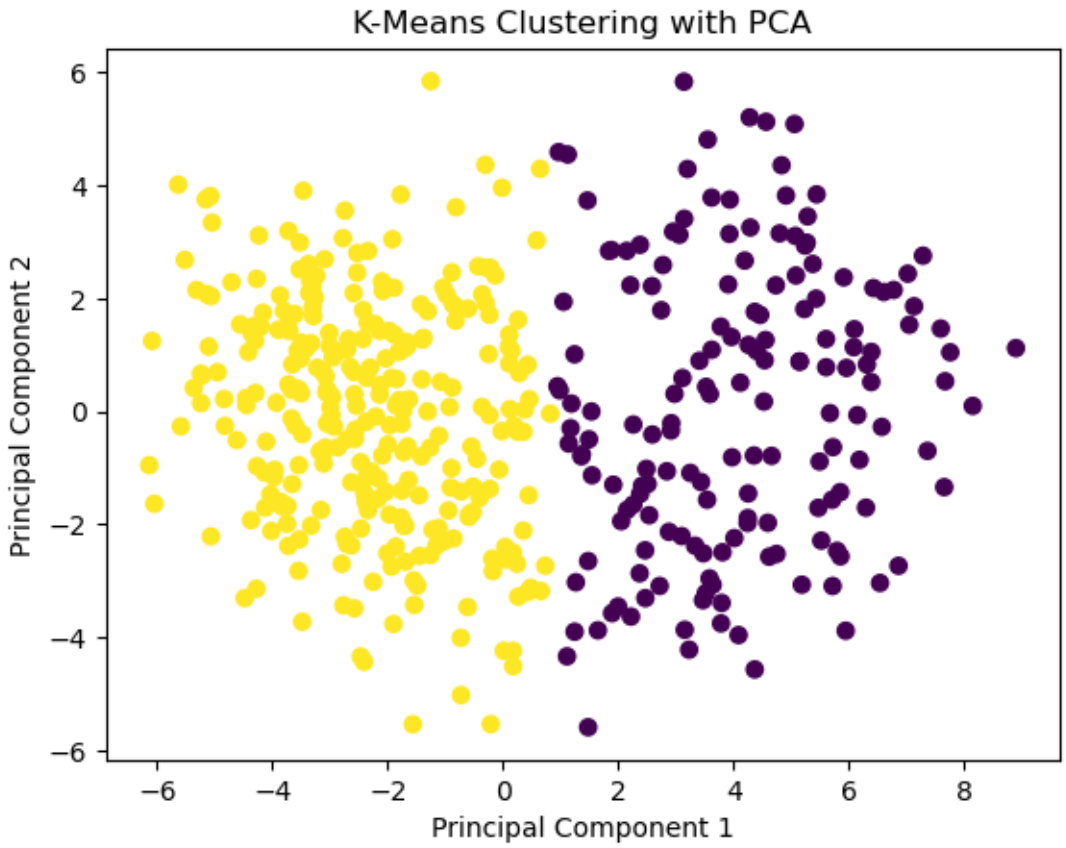


[ ]:
```
kmeansFinal = KMeans(n_clusters=2)
kmeansFinal.fit(X_train,y_train)
```

[ ]:
```
KMeans(n_clusters=2)
```

[ ]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)  # Reducing to 2 dimensions
X_train_pca = pca.fit_transform(X_train)
```

[ ]:
```
# Get the cluster assignments
labels = kmeansFinal.labels_

# Plot the data points with different colors for each cluster
plt.scatter(X_train_pca[:, 0], X_train_pca[:, 1], c=labels)
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-Means Clustering with PCA')
plt.show()
```



9 Conclusion:

So from both elbow method and silhouette coefficient analysis it is clear that the no of cluste