Practical_5

July 16, 2025

1 Practical 5

220107089 | Altynbek Adilkhan

Import the modules required

[17]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix
```

Read the file "cancer.csv" and show the first 5 rows

[18]:
```
CSV_FILE = pd.read_csv('Desktop/SDU/DBMS2/cancer.csv')
CSV_FILE.head(5)
```

[18]:
```
         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

   …  texture_worst  perimeter_worst  area_worst  smoothness_worst  \
0  …          17.33           184.60      2019.0            0.1622
1  …          23.41           158.80      1956.0            0.1238
2  …          25.53           152.50      1709.0            0.1444
3  …          26.50            98.87       567.7            0.2098
4  …          16.67           152.20      1575.0            0.1374

   compactness_worst  concavity_worst  concave points_worst  symmetry_worst  \
0             0.6656           0.7119                0.2654          0.4601
```

```
1             0.1866            0.2416                  0.1860          0.2750
2             0.4245            0.4504                  0.2430          0.3613
3             0.8663            0.6869                  0.2575          0.6638
4             0.2050            0.4000                  0.1625          0.2364

   fractal_dimension_worst  Unnamed: 32
0                  0.11890          NaN
1                  0.08902          NaN
2                  0.08758          NaN
3                  0.17300          NaN
4                  0.07678          NaN

[5 rows x 33 columns]
```

1.0.1 Q1: Group the diagnosis by radius area and add "value_accounts()" method to show the counts.

Hint: check the following documentations for the functions that you will use. Group the diagnosis by the radius area

```
[19]: grouped_counts = CSV_FILE.groupby('radius_mean')['diagnosis'].value_counts()
grouped_counts_df = grouped_counts.reset_index(name='count')
grouped_counts_df.head()
```

```
[19]:    radius_mean diagnosis  count
      0        6.981         B      1
      1        7.691         B      1
      2        7.729         B      1
      3        7.760         B      1
      4        8.196         B      1
```

1.0.2 Q2: Explain what did you get.

In this problem, we grouped the dataset by radius_mean (the average radius of cancer cells in the dataset) and then counted the number of malignant (M) and benign (B) diagnoses for each radius_mean value.

The DataFrame contains: For each unique radius_mean value, we have the number of benign and malignant tumor diagnoses.

Meaning: The output shows the distribution of diagnoses (benign and malignant) for different values of radius_mean, which helps us understand how the radius of cancer cells may correlate with the type of diagnosis.

1.0.3 Q3: Use DataFram method "crosstab()" to apply cross tabulation between diagnosis and radius mean

Get this intersting data in a table form. Use crosstab

[20]:

```
crosstab_result = pd.crosstab(CSV_FILE['radius_mean'], CSV_FILE['diagnosis'])
crosstab_result.head()
```

[20]:

```
diagnosis    B  M
radius_mean
6.981        1  0
7.691        1  0
7.729        1  0
7.760        1  0
8.196        1  0
```

1.0.4 Q4: Check the doccumentation of drop function and do the following:

```
Q4_1: Drop only id column
    #drop id column
    #show the first 5 rows after dropinng id
Q4_2: use only one command to drop columns 7 up to the last one
    #use only one command to drop columns 7 up to the last one
    #show the first 5 rows after dropinng id
```

Q4_1: Drop only id column

```
    #drop id column
    #show the first 5 rows after dropinng id
```

[21]:

```
df_q4_1 = CSV_FILE.drop('id', axis=1)
df_q4_1.head()
```

[21]:

```
  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  …  texture_worst  perimeter_worst  area_worst  \
0         0.2419  …          17.33           184.60      2019.0
1         0.1812  …          23.41           158.80      1956.0
2         0.2069  …          25.53           152.50      1709.0
3         0.2597  …          26.50            98.87       567.7
4         0.1809  …          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  Unnamed: 32
0          0.4601                  0.11890          NaN
1          0.2750                  0.08902          NaN
2          0.3613                  0.08758          NaN
3          0.6638                  0.17300          NaN
4          0.2364                  0.07678          NaN

[5 rows x 32 columns]
```

Q4_2: use only one command to drop columns 7 up to the last one

```
#use only one command to drop columns 7 up to the last one
#show the first 5 rows after dropinng id
```

[22]:
```
df_q4_2 = CSV_FILE.drop(CSV_FILE.columns[7:], axis=1)
df_q4_2.head()
```

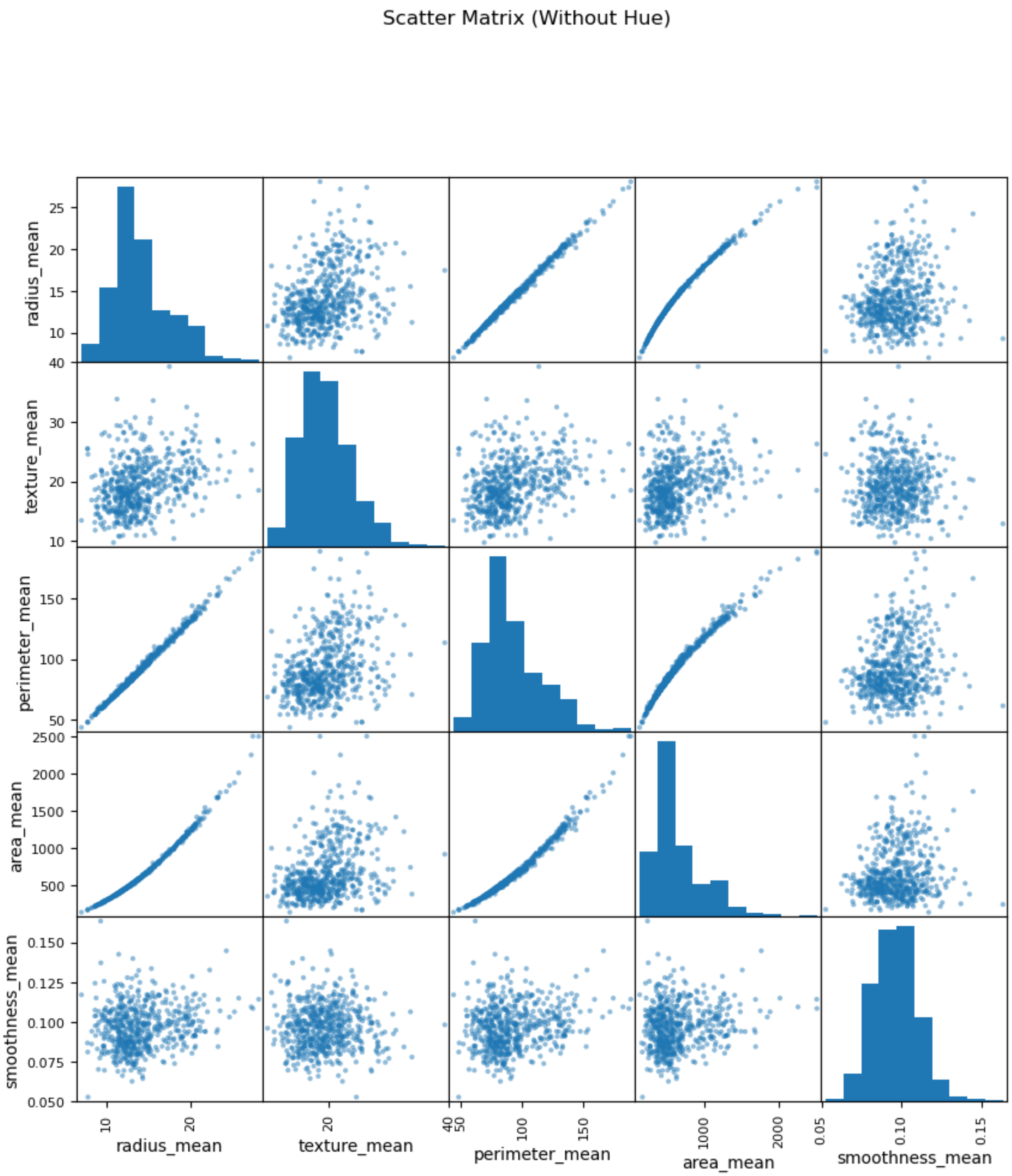
[22]:
```
         id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0    842302         M        17.99         10.38          122.80     1001.0
1    842517         M        20.57         17.77          132.90     1326.0
2  84300903         M        19.69         21.25          130.00     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
0          0.11840
1          0.08474
2          0.10960
3          0.14250
4          0.10030
```

1.0.5 Q5: Draw a scatter matrix using seaborn. Make sure that you finish question Q4 first.

```
#Draw scatter martix using seaborn
```

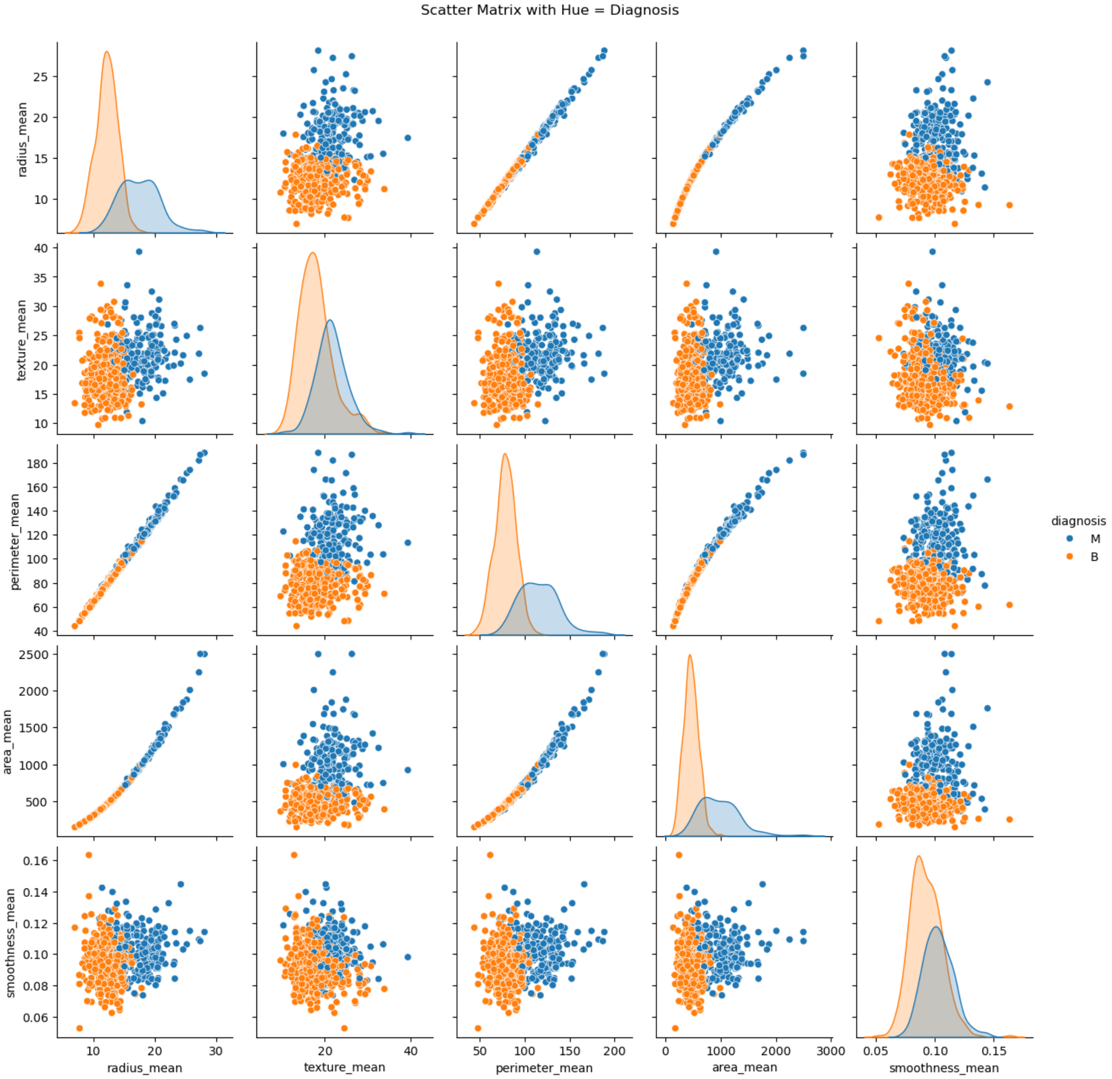
[23]:
```
scatter_matrix(df_q4_1.iloc[:, 1:6], figsize=(10, 10))
plt.suptitle("Scatter Matrix (Without Hue)", y=1.02)
plt.show()
```



1.0.6 Q6: Draw a scatter matrix using seaborn. Add hue argument (Do you know which variable should be in hue argument?)

#Draw scatter martix using seaborn

[24]:
```
sns.pairplot(df_q4_1.iloc[:, 1:6].join(df_q4_1['diagnosis']), hue='diagnosis', diag_kind='kde')
plt.suptitle("Scatter Matrix with Hue = Diagnosis", y=1.02)
plt.show()
```



1.0.7 Q7: calculate the correlation matrix and print it.

```
#calculate correlaion matrix. use numeric_only argument inside the correleation functi
#show it
```

[25]:

```
correlation_matrix = df_q4_1.corr(numeric_only=True)
correlation_matrix
```

[25]:

```
                         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
Unnamed: 32                    NaN          NaN              NaN        NaN

                         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
Unnamed: 32                          NaN               NaN             NaN

                         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
Unnamed: 32                              NaN            NaN
```

```
                         fractal_dimension_mean  …  texture_worst  \
radius_mean                           -0.311631  …       0.297008
texture_mean                          -0.076437  …       0.912045
perimeter_mean                        -0.261477  …       0.303038
area_mean                             -0.283110  …       0.287489
smoothness_mean                        0.584792  …       0.036072
compactness_mean                       0.565369  …       0.248133
concavity_mean                         0.336783  …       0.299879
concave points_mean                    0.166917  …       0.292752
symmetry_mean                          0.479921  …       0.090651
fractal_dimension_mean                 1.000000  …      -0.051269
radius_se                              0.000111  …       0.194799
texture_se                             0.164174  …       0.409003
perimeter_se                           0.039830  …       0.200371
area_se                               -0.090170  …       0.196497
smoothness_se                          0.401964  …      -0.074743
compactness_se                         0.559837  …       0.143003
concavity_se                           0.446630  …       0.100241
concave points_se                      0.341198  …       0.086741
symmetry_se                            0.345007  …      -0.077473
fractal_dimension_se                   0.688132  …      -0.003195
radius_worst                          -0.253691  …       0.359921
texture_worst                         -0.051269  …       1.000000
perimeter_worst                       -0.205151  …       0.365098
area_worst                            -0.231854  …       0.345842
smoothness_worst                       0.504942  …       0.225429
compactness_worst                      0.458798  …       0.360832
concavity_worst                        0.346234  …       0.368366
concave points_worst                   0.175325  …       0.359755
symmetry_worst                         0.334019  …       0.233027
fractal_dimension_worst                0.767297  …       0.219122
Unnamed: 32                                 NaN  …            NaN

                         perimeter_worst  area_worst  smoothness_worst  \
radius_mean                     0.965137    0.941082          0.119616
texture_mean                    0.358040    0.343546          0.077503
perimeter_mean                  0.970387    0.941550          0.150549
area_mean                       0.959120    0.959213          0.123523
smoothness_mean                 0.238853    0.206718          0.805324
compactness_mean                0.590210    0.509604          0.565541
concavity_mean                  0.729565    0.675987          0.448822
concave points_mean             0.855923    0.809630          0.452753
symmetry_mean                   0.219169    0.177193          0.426675
fractal_dimension_mean         -0.205151   -0.231854          0.504942
radius_se                       0.719684    0.751548          0.141919
texture_se                     -0.102242   -0.083195         -0.073658
```

```
perimeter_se                    0.721031    0.730713          0.130054
area_se                         0.761213    0.811408          0.125389
smoothness_se                  -0.217304   -0.182195          0.314457
compactness_se                  0.260516    0.199371          0.227394
concavity_se                    0.226680    0.188353          0.168481
concave points_se               0.394999    0.342271          0.215351
symmetry_se                    -0.103753   -0.110343         -0.012662
fractal_dimension_se           -0.001000   -0.022736          0.170568
radius_worst                    0.993708    0.984015          0.216574
texture_worst                   0.365098    0.345842          0.225429
perimeter_worst                 1.000000    0.977578          0.236775
area_worst                      0.977578    1.000000          0.209145
smoothness_worst                0.236775    0.209145          1.000000
compactness_worst               0.529408    0.438296          0.568187
concavity_worst                 0.618344    0.543331          0.518523
concave points_worst            0.816322    0.747419          0.547691
symmetry_worst                  0.269493    0.209146          0.493838
fractal_dimension_worst         0.138957    0.079647          0.617624
Unnamed: 32                          NaN         NaN               NaN

                         compactness_worst  concavity_worst  \
radius_mean                       0.413463         0.526911
texture_mean                      0.277830         0.301025
perimeter_mean                    0.455774         0.563879
area_mean                         0.390410         0.512606
smoothness_mean                   0.472468         0.434926
compactness_mean                  0.865809         0.816275
concavity_mean                    0.754968         0.884103
concave points_mean               0.667454         0.752399
symmetry_mean                     0.473200         0.433721
fractal_dimension_mean            0.458798         0.346234
radius_se                         0.287103         0.380585
texture_se                       -0.092439        -0.068956
perimeter_se                      0.341919         0.418899
area_se                           0.283257         0.385100
smoothness_se                    -0.055558        -0.058298
compactness_se                    0.678780         0.639147
concavity_se                      0.484858         0.662564
concave points_se                 0.452888         0.549592
symmetry_se                       0.060255         0.037119
fractal_dimension_se              0.390159         0.379975
radius_worst                      0.475820         0.573975
texture_worst                     0.360832         0.368366
perimeter_worst                   0.529408         0.618344
area_worst                        0.438296         0.543331
smoothness_worst                  0.568187         0.518523
compactness_worst                 1.000000         0.892261
```

```
concavity_worst                   0.892261         1.000000
concave points_worst              0.801080         0.855434
symmetry_worst                    0.614441         0.532520
fractal_dimension_worst           0.810455         0.686511
Unnamed: 32                            NaN              NaN

                         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
Unnamed: 32                               NaN             NaN

                         fractal_dimension_worst  Unnamed: 32
radius_mean                             0.007066          NaN
texture_mean                            0.119205          NaN
perimeter_mean                          0.051019          NaN
area_mean                               0.003738          NaN
smoothness_mean                         0.499316          NaN
compactness_mean                        0.687382          NaN
concavity_mean                          0.514930          NaN
```

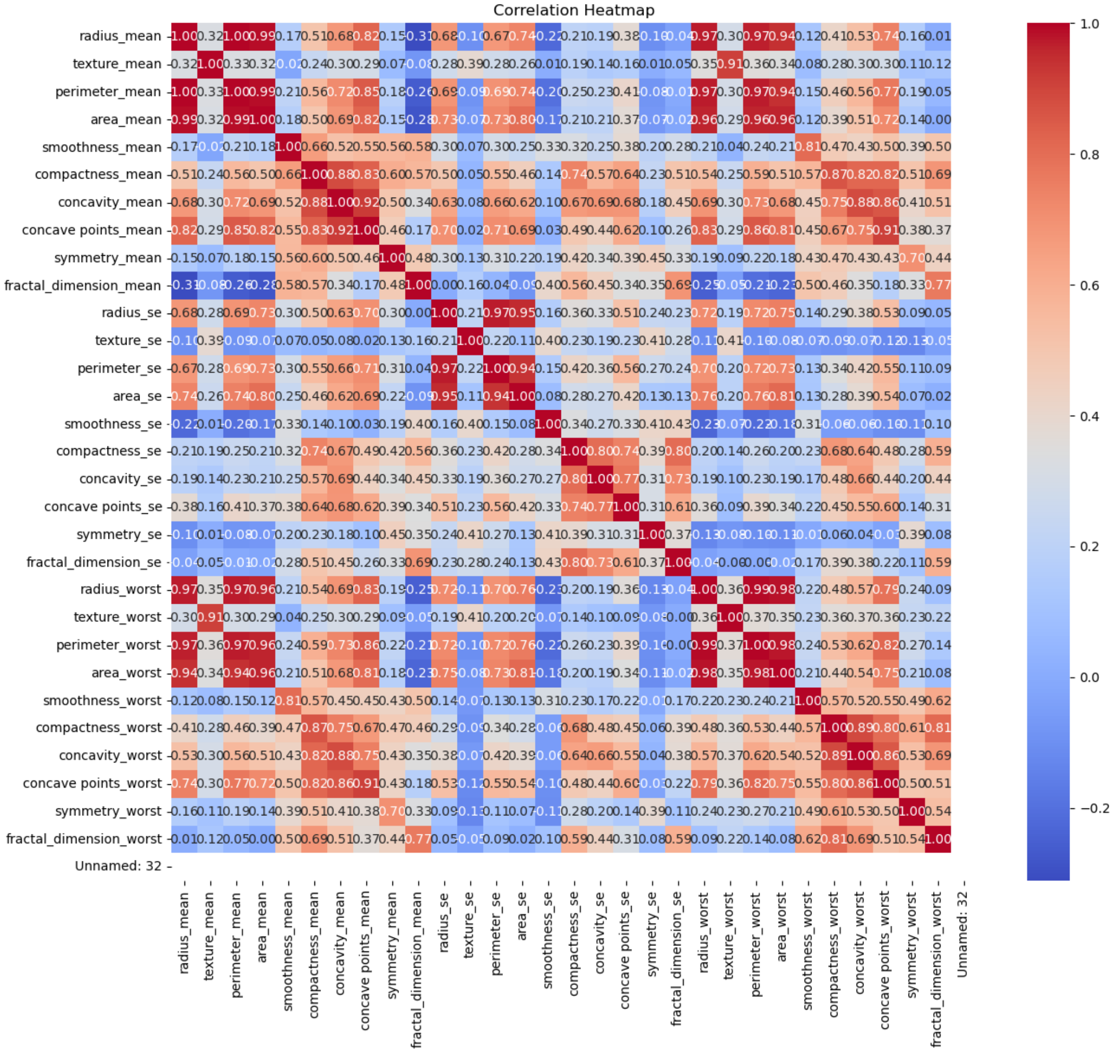
```
concave points_mean                     0.368661          NaN
symmetry_mean                           0.438413          NaN
fractal_dimension_mean                  0.767297          NaN
radius_se                               0.049559          NaN
texture_se                             -0.045655          NaN
perimeter_se                            0.085433          NaN
area_se                                 0.017539          NaN
smoothness_se                           0.101480          NaN
compactness_se                          0.590973          NaN
concavity_se                            0.439329          NaN
concave points_se                       0.310655          NaN
symmetry_se                             0.078079          NaN
fractal_dimension_se                    0.591328          NaN
radius_worst                            0.093492          NaN
texture_worst                           0.219122          NaN
perimeter_worst                         0.138957          NaN
area_worst                              0.079647          NaN
smoothness_worst                        0.617624          NaN
compactness_worst                       0.810455          NaN
concavity_worst                         0.686511          NaN
concave points_worst                    0.511114          NaN
symmetry_worst                          0.537848          NaN
fractal_dimension_worst                 1.000000          NaN
Unnamed: 32                                  NaN          NaN

[31 rows x 31 columns]
```

1.0.8 Q8: Draw a heat map for your dataset. Don't forget to resize the figure with appropriate sizing values.

#Change the color map, you need to use cmap options

```
[26]: plt.figure(figsize=(14, 12))
      sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
      plt.title("Correlation Heatmap")
      plt.show()
```


Correlation Heatmap
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
fractal_dimension_worst
Unnamed: 32
1.0
0.8
0.6
0.4
0.2
0.0
−0.2