ICP -1: SCREENSHOTS OF CODE AND OUTPUT

- ICP 1

Perform three data analysis tasks on the data using google colab

▼ Importing Libraries and data

[11] import pandas as pd # data processing, reading CSV file import seaborn as sns # data visualization library import matplotlib.pyplot as plt

/usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions in the public API at pandas.testing instead. import pandas.util.testing as tm

[3] path = '/content/drive/My Drive/Data.csv'
df= pd.read_csv(path)

[4] df.head() # head method shows first 5 rows

₽ concave id diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean symmetry_mean fractal_dimension_mean radius points_mean 842302 M 17.99 10.38 122.80 1001.0 0.11840 0.27760 0.3001 0.14710 0.2419 0.07871 1.0 842517 M 20.57 17.77 132.90 1326.0 0.08474 0.07864 0.0869 0.07017 0.1812 0.05667 0.5 2 84300903 0.15990 М 19.69 21.25 130.00 1203.0 0.10960 0.1974 0.12790 0.2069 0.05999 0.7 3 84348301 20.38 386.1 0.14250 0.28390 0.2414 0.10520 0.2597 0.09744 M 11.42 77.58 0.4 4 84358402 20.29 14.34 135.10 1297.0 0.10030 0.13280 0.1980 0.10430 0.1809 0.05883

▼ 1. Descriptive Statistics

[8] df.info() # to view data type, NA values and no.of rows and columns

C <class 'pandas.core.frame.DataFrame'>
 RangeIndex: 569 entries, 0 to 568
 Data columns (total 33 columns):

#	Column	Non-Null Count	Dtype									
0	id	569 non-null	int64									
1	diagnosis	569 non-null	object									
2	radius_mean	569 non-null	float64									
3	texture_mean	569 non-null	float64									
4	perimeter_mean	569 non-null	float64									
5	area_mean	569 non-null	float64									
6	smoothness_mean	569 non-null	float64									
7	compactness_mean	569 non-null	float64									
8	concavity_mean	569 non-null	float64									
9	concave points_mean	569 non-null	float64									
10	symmetry_mean	569 non-null	float64									
11	fractal_dimension_mean	569 non-null	float64									
12	radius_se	569 non-null	float64									
13	texture_se	569 non-null	float64									
14	perimeter_se	569 non-null	float64									
15	area_se	569 non-null	float64									
16	smoothness_se	569 non-null	float64									
17	compactness_se	569 non-null	float64									
18	concavity_se	569 non-null	float64									
19	concave points_se	569 non-null	float64									
20	symmetry_se	569 non-null	float64									
21	fractal_dimension_se	569 non-null	float64									
22	radius_worst	569 non-null	float64									
23	texture_worst	569 non-null	float64									
24	perimeter_worst	569 non-null	float64									
25	area_worst	569 non-null	float64									
26	smoothness_worst	569 non-null	float64									
27	compactness_worst	569 non-null	float64									
28	concavity_worst	569 non-null	float64									
29	concave points_worst	569 non-null	float64									
30	symmetry_worst	569 non-null	float64									
31	fractal_dimension_worst	569 non-null	float64									
32	Unnamed: 32	0 non-null	float64									
	dtypes: float64(31), int64(1), object(1)											

memory usage: 146.8+ KB

₽

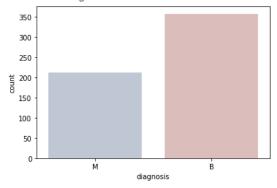
>

J

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fr
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	

ax = sns.countplot(df.diagnosis,label="Count", palette="vlag")
B, M = df.diagnosis.value_counts()
print('Number of Benign: ',B)
print('Number of Malignant : ',M)

Number of Benign: 357
Number of Malignant: 212



▼ 2. HEATMAP

To find correlation of all the quantitative variables.

radius_mean ·	1.0	0.3	1.0	1.0	0.2	0.5	0.7	0.8	0.1	-0.3	0.7	-0.1	0.7	0.7	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	0.9	0.1	0.4	0.5	0.7	0.2	0.0
texture_mean -	0.3	1.0	0.3	0.3	-0.0	0.2	0.3	0.3	0.1	-0.1	0.3	0.4	0.3	0.3	0.0	0.2	0.1	0.2	0.0	0.1	0.4	0.9	0.4	0.3	0.1	0.3	0.3	0.3	0.1	0.1
perimeter_mean ·	1.0	0.3	1.0	1.0	0.2	0.6	0.7	0.9	0.2	-0.3	0.7	-0.1	0.7	0.7	-0.2	0.3	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	0.9	0.2	0.5	0.6	0.8	0.2	0.1
area_mean -	1.0	0.3	1.0	1.0	0.2	0.5	0.7	0.8	0.2	-0.3	0.7	-0.1	0.7	0.8	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	1.0	0.1	0.4	0.5	0.7	0.1	0.0
smoothness_mean -	0.2	-0.0	0.2	0.2	1.0	0.7	0.5	0.6	0.6	0.6	0.3	0.1	0.3	0.2	0.3	0.3	0.2	0.4	0.2	0.3	0.2	0.0	0.2	0.2	0.8	0.5	0.4	0.5	0.4	0.5
compactness_mean -	0.5	0.2	0.6	0.5	0.7	1.0	0.9	0.8	0.6	0.6	0.5	0.0	0.5	0.5	0.1	0.7	0.6	0.6	0.2	0.5	0.5	0.2	0.6	0.5	0.6	0.9	0.8	0.8	0.5	0.7
concavity_mean	0.7	0.3	0.7	0.7	0.5	0.9	1.0	0.9	0.5	0.3	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.7	0.2	0.4	0.7	0.3	0.7	0.7	0.4	0.8	0.9	0.9	0.4	0.5
concave points_mean	0.8	0.3	0.9	0.8	0.6	0.8	0.9	1.0	0.5	0.2	0.7	0.0	0.7	0.7	0.0	0.5	0.4	0.6	0.1	0.3	0.8	0.3	0.9	0.8	0.5	0.7	0.8	0.9	0.4	0.4
symmetry_mean -	0.1	0.1	0.2	0.2	0.6	0.6	0.5	0.5	1.0	0.5	0.3	0.1	0.3	0.2	0.2	0.4	0.3	0.4	0.4		0.2	0.1	0.2	0.2	0.4	0.5	0.4	0.4	0.7	0.4
fractal_dimension_mean -	-0.3	-0.1	-0.3	-0.3	0.6	0.6	0.3	0.2	0.5	1.0	0.0	0.2	0.0	-0.1	0.4	0.6	0.4	0.3	0.3	0.7	-0.3	-0.1	-0.2	-0.2	0.5	0.5	0.3	0.2	0.3	0.8
radius_se	0.7	0.3	0.7	0.7	0.3	0.5	0.6	0.7	0.3	0.0	1.0	0.2	1.0	1.0	0.2	0.4	0.3	0.5	0.2	0.2	0.7	0.2	0.7	0.8	0.1	0.3	0.4	0.5	0.1	0.0
texture_se -	-0.1	0.4	-0.1	-0.1	0.1	0.0	0.1	0.0	0.1	0.2	0.2	1.0	0.2	0.1	0.4	0.2	0.2	0.2	0.4	0.3	-0.1	0.4	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0
perimeter_se	0.7	0.3	0.7	0.7	0.3	0.5	0.7	0.7	0.3	0.0	1.0	0.2	1.0	0.9	0.2	0.4	0.4	0.6	0.3	0.2	0.7	0.2	0.7	0.7	0.1	0.3	0.4	0.6	0.1	0.1
area_se -	0.7	0.3	0.7	0.8	0.2	0.5	0.6	0.7	0.2	-0.1	1.0	0.1	0.9	1.0	0.1	0.3	0.3	0.4	0.1	0.1	0.8	0.2	0.8	0.8	0.1	0.3	0.4	0.5	0.1	0.0
smoothness_se	-0.2	0.0	-0.2	-0.2	0.3	0.1	0.1	0.0	0.2	0.4	0.2	0.4	0.2	0.1	1.0	0.3	0.3	0.3	0.4	0.4	-0.2	-0.1	-0.2	-0.2	0.3	-0.1	-0.1	-0.1	-0.1	0.1
compactness_se -	0.2	0.2	0.3	0.2	0.3	0.7	0.7	0.5	0.4	0.6	0.4	0.2	0.4	0.3	0.3	1.0	0.8	0.7	0.4	0.8	0.2	0.1	0.3	0.2	0.2	0.7	0.6	0.5	0.3	0.6
concavity_se	0.2	0.1	0.2	0.2	0.2	0.6	0.7	0.4	0.3	0.4	0.3	0.2	0.4	0.3	0.3	0.8	1.0	0.8	0.3	0.7	0.2	0.1	0.2	0.2	0.2	0.5	0.7	0.4	0.2	0.4
concave points_se	0.4	0.2	0.4	0.4	0.4	0.6	0.7	0.6	0.4	0.3	0.5	0.2	0.6	0.4	0.3	0.7	0.8	1.0	0.3	0.6	0.4	0.1	0.4	0.3	0.2	0.5	0.5	0.6	0.1	0.3
symmetry_se	-0.1	0.0	-0.1	-0.1	0.2	0.2	0.2	0.1	0.4	0.3	0.2	0.4	0.3	0.1		0.4	0.3	0.3	1.0	0.4	-0.1	-0.1	-0.1	-0.1	-0.0	0.1	0.0	-0.0	0.4	0.1
fractal_dimension_se	-0.0	0.1	-0.0	-0.0	0.3	0.5	0.4	0.3	0.3	0.7	0.2	0.3	0.2	0.1	0.4	0.8	0.7	0.6	0.4	1.0	-0.0	-0.0	-0.0	-0.0	0.2	0.4	0.4	0.2	0.1	0.6
radius_worst -	1.0	0.4	1.0	1.0	0.2	0.5	0.7	0.8	0.2	-0.3	0.7	-0.1	0.7	0.8	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.4	1.0	1.0	0.2	0.5	0.6	0.8	0.2	0.1
texture_worst -	0.3	0.9	0.3	0.3	0.0	0.2	0.3	0.3	0.1	-0.1	0.2	0.4	0.2	0.2	-0.1	0.1	0.1	0.1	-0.1	-0.0	0.4	1.0	0.4	0.3	0.2	0.4	0.4	0.4	0.2	0.2
perimeter_worst	1.0	0.4	1.0	1.0	0.2	0.6	0.7	0.9	0.2	-0.2	0.7	-0.1	0.7	0.8	-0.2	0.3	0.2	0.4	-0.1	-0.0	1.0	0.4	1.0	1.0	0.2	0.5	0.6	0.8	0.3	0.1
area_worst -	0.9	0.3	0.9	1.0	0.2	0.5	0.7	0.8	0.2	-0.2	0.8	-0.1	0.7	0.8	-0.2	0.2	0.2	0.3	-0.1	-0.0	1.0	0.3	1.0	1.0	0.2	0.4	0.5	0.7	0.2	0.1
smoothness_worst	0.1	0.1	0.2	0.1	0.8	0.6	0.4	0.5	0.4	0.5	0.1	-0.1	0.1	0.1	0.3	0.2	0.2	0.2	-0.0	0.2	0.2	0.2	0.2	0.2	1.0	0.6	0.5	0.5	0.5	0.6
compactness_worst -	0.4	0.3	0.5	0.4	0.5	0.9	0.8	0.7	0.5	0.5	0.3	-0.1	0.3	0.3	-0.1	0.7	0.5	0.5	0.1	0.4	0.5	0.4	0.5	0.4	0.6	1.0	0.9	0.8	0.6	0.8
concavity_worst	0.5	0.3	0.6	0.5	0.4	0.8	0.9	0.8	0.4	0.3	0.4	-0.1	0.4	0.4	-0.1	0.6	0.7	0.5	0.0	0.4	0.6	0.4	0.6	0.5	0.5	0.9	1.0	0.9	0.5	0.7
concave points_worst	0.7	0.3	0.8	0.7	0.5	0.8	0.9	0.9	0.4	0.2	0.5	-0.1	0.6	0.5	-0.1	0.5	0.4	0.6	-0.0	0.2	0.8	0.4	0.8	0.7	0.5	0.8	0.9	1.0	0.5	0.5
symmetry_worst -	0.2	0.1	0.2	0.1	0.4	0.5	0.4	0.4	0.7	0.3	0.1	-0.1	0.1	0.1	-0.1	0.3	0.2	0.1	0.4	0.1	0.2	0.2	0.3	0.2	0.5	0.6	0.5	0.5	1.0	0.5
fractal_dimension_worst	0.0	0.1	0.1	0.0	0.5	0.7	0.5	0.4	0.4	0.8	0.0	-0.0	0.1	0.0	0.1	0.6	0.4	0.3	0.1	0.6	0.1	0.2	0.1	0.1	0.6	0.8	0.7	0.5	0.5	1.0
	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

1.0

- 0.8

- 0.6

- 0.4

- 0.2

- 0.0

- -0.2

3. Boxplots to identify important features

```
data = pd.melt(df,id_vars="diagnosis",
                                                                                                                               var name="features",
                                                                                                                               value_name='value')
                       plt.figure(figsize=(10,10))
                       sns.boxplot(x="features", y="value", hue="diagnosis", data=data)
                       plt.xticks(rotation=90)
  (array([ 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]),
                            <a list of 32 Text major ticklabel objects>)
                                                                                                                                                                                                                                                                                                                                                                                                  diagnosis
                                                                                                                                                                                                                                                                                                                                                                                                    В В
                                             radius_mean
rexture_mean
area_mean
area_mean
area_mean
concavity_mean
symmetry_mean
radius_se
rexture_se
area_se
area_
```

```
There are lots of outliers as the data is not standardized. And there are lots of predictors so the boxplots are not visible
 outcome = df.diagnosis
                                                   # outcome variable is diagnosis
     list = ['Unnamed: 32','id','diagnosis']
      predictors = df.drop(list,axis = 1 ) # the remaining variables are designated as predictors
     data = (predictors - predictors.mean()) / (predictors.std()) # To standardize data
     data = pd.concat([outcome,data.iloc[:,0:10]],axis=1) # Limiting to 10 predictors
     data = pd.melt(data,id_vars="diagnosis",
                        var_name="features",
                        value_name='value')
      plt.figure(figsize=(10,10))
      sns.boxplot(x="features", y="value", hue="diagnosis", data=data, palette="rocket")
      plt.xticks(rotation=90)
 ₽
                diagnosis
\supset
```

Malignant cells have a higher value for all the 10 predictors when compared to benign cells