

DEVP PROJECT-1

Analyzing the Breast Cancer Wisconsin Dataset using Python to extract various insights that would help various organizations and the society.

Vasu Bansal

045055

PGDM-BDA (04)

Project Objective:

To conduct a comprehensive analysis of this dataset to find hidden trends that may help in the early detection of breast cancer among patients by using techniques such as correlation between various features which may indicate an early onset. This analysis would help doctors identify the patients in their early stages of cancer and improve the chances of patient survival with minimum side effects from the disease and its treatment.

General Description of Dataset:

	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	
..	
564	926424	M	21.56	22.39	142.00	1479.0	
565	926682	M	20.13	28.25	131.20	1261.0	
566	926954	M	16.60	28.08	108.30	858.1	
567	927241	M	20.60	29.33	140.10	1265.0	
568	92751	B	7.76	24.54	47.92	181.0	
	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\		
0	0.11840	0.27760	0.30010	0.14710			
1	0.08474	0.07864	0.08690	0.07017			
2	0.10960	0.15990	0.19740	0.12790			
3	0.14250	0.28390	0.24140	0.10520			
4	0.10030	0.13280	0.19800	0.10430			
..			
564	0.11100	0.11590	0.24390	0.13890			
565	0.09780	0.10340	0.14400	0.09791			
566	0.08455	0.10230	0.09251	0.05302			
567	0.11780	0.27700	0.35140	0.15200			
568	0.05263	0.04362	0.00000	0.00000			
	texture_worst	perimeter_worst	area_worst	smoothness_worst	\		
0	...	17.33	184.60	2019.0	0.16220		
1	...	23.41	158.80	1956.0	0.12380		
2	...	25.53	152.50	1709.0	0.14440		
3	...	26.50	98.87	567.7	0.20980		
4	...	16.67	152.20	1575.0	0.13740		
..		
564	...	26.40	166.10	2027.0	0.14100		
565	...	38.25	155.00	1731.0	0.11660		

```

▶ .. ...
564 ... 26.40 166.10 2027.0 0.14100
565 ... 38.25 155.00 1731.0 0.11660
566 ... 34.12 126.70 1124.0 0.11390
567 ... 39.42 184.60 1821.0 0.16500
568 ... 30.37 59.16 268.6 0.08996

compactness_worst concavity_worst concave points_worst symmetry_worst \
0 0.66560 0.7119 0.2654 0.4601
1 0.18660 0.2416 0.1860 0.2750
2 0.42450 0.4504 0.2430 0.3613
3 0.86630 0.6869 0.2575 0.6638
4 0.20500 0.4000 0.1625 0.2364
.. ...
564 0.21130 0.4107 0.2216 0.2060
565 0.19220 0.3215 0.1628 0.2572
566 0.30940 0.3403 0.1418 0.2218
567 0.86810 0.9387 0.2650 0.4087
568 0.06444 0.0000 0.0000 0.2871

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
.. ...
564 0.07115 NaN
565 0.06637 NaN
566 0.07820 NaN
567 0.12400 NaN
568 0.07039 NaN

[569 rows x 33 columns]

```

This data frame consists of information about 569 patients, and there are 30 factors based on which a patient is told if their cancer is Malignant (non-curable) [M] or Benign(curable)[B].

Column description-

1. ID: This column contains unique identification numbers for each patient. It's used for tracking and reference purposes but doesn't provide any meaningful information for analysis so we will be dropping this column.

2. Diagnosis: This is the main target variable. It represents the diagnosis of the breast tumor:

'M' indicates a malignant tumor, which means it is cancerous.

'B' indicates a benign tumor, which means it is non-cancerous.

3. Mean Radius: This feature represents the mean distance from the center to points on the perimeter of the tumor mass. It gives an idea of the average size of the tumor.

4. Mean Texture: It represents the mean value of the gray-scale texture of the pixels in the tumor mass. Texture refers to patterns in pixel intensities, and this feature measures the average texture.

5. Mean Perimeter: This column represents the mean perimeter or circumference of the tumor mass. It provides information about the average outline of the tumor.

6. Mean Area: It represents the mean area of the tumor mass. This feature quantifies the average size of the tumor in terms of the number of pixels it covers.

7. Mean Smoothness: This feature represents the mean of local variation in radius lengths. It characterizes how smooth or irregular the boundaries of the tumor are.

8. Mean Compactness: It represents the mean of the compactness of the tumor mass. Compactness is calculated as $(\text{perimeter}^2 / \text{area} - 1.0)$. It describes how closely the tumor mass resembles a perfect circle.

9. Mean Concavity: This column represents the mean severity of concave portions of the contour of the tumor mass. It measures the depth and severity of inwardly curved portions of the tumor boundary.

10. Mean Concave Points: It represents the mean number of concave portions of the contour. It quantifies how many concave regions are present in the tumor.

11. Mean Symmetry: This feature represents the mean symmetry of the tumor mass. It measures how symmetrical or asymmetrical the tumor is in terms of shape.

12. Mean Fractal Dimension: It represents the mean "coastline approximation" of the tumor mass. Fractal dimension characterizes the complexity of the tumor boundary.

13. Radius SE: This column represents the standard error of the mean radius. It indicates the variability or uncertainty in the mean radius measurement.

14. Texture SE: It represents the standard error of the mean texture. It measures the variability or uncertainty in the mean texture value.

15. Perimeter SE: This column represents the standard error of the mean perimeter. It quantifies the variability or uncertainty in the mean perimeter measurement.

16. Area SE: It represents the standard error of the mean area. It indicates the variability or uncertainty in the mean area measurement.

17. Smoothness SE: This feature represents the standard error of the mean smoothness. It measures the variability or uncertainty in the mean smoothness value.

18. Compactness SE: It represents the standard error of the mean compactness. It quantifies the variability or uncertainty in the mean compactness value.

19. Concavity SE: This column represents the standard error of the mean concavity. It indicates the variability or uncertainty in the mean concavity measurement.

20. Concave Points SE: It represents the standard error of the mean number of concave portions of the contour. It measures the variability in this aspect of the tumor.

21. Symmetry SE: This feature represents the standard error of the mean symmetry. It quantifies the variability or uncertainty in the mean symmetry value.

22. Fractal Dimension SE: It represents the standard error of the mean fractal dimension. It indicates the variability or uncertainty in the mean fractal dimension measurement.

23. Worst Radius: This column represents the "worst" or largest mean value of the radius in the tumor mass. It provides information about the largest size observed in the tumor.

24. Worst Texture: It represents the "worst" or largest mean value of the texture in the tumor mass. It characterizes the highest observed texture.

25. Worst Perimeter: This column represents the "worst" or largest mean value of the perimeter in the tumor mass. It quantifies the largest observed perimeter.

26. Worst Area: It represents the "worst" or largest mean value of the area in the tumor mass. It quantifies the largest observed area.

27. Worst Smoothness: This feature represents the "worst" or largest mean value of local variation in radius lengths. It characterizes the highest observed smoothness variability.

28. Worst Compactness: It represents the "worst" or largest mean value of the compactness of the tumor mass. It quantifies the highest observed compactness.


29. Worst Concavity: This column represents the "worst" or largest mean value of the severity of concave portions of the contour. It characterizes the highest observed concavity.


30. Worst Concave Points: It represents the "worst" or largest mean value of the number of concave portions of the contour. It quantifies the highest observed number of concave regions.



31. Worst Symmetry: This feature represents the "worst" or largest mean value of symmetry in the tumor mass. It characterizes the highest observed symmetry.

32. Worst Fractal Dimension: It represents the "worst" or largest mean value of the fractal dimension in the tumor mass. It quantifies the highest observed fractal complexity.

These columns provide a comprehensive set of measurements and characteristics of breast cancer tumors, which can be used for analysis, classification, and gaining insights into the nature of these tumors, aiding in diagnosis and treatment decisions.

 #dataset description
df.describe().T

		count	mean	std	min	25%	50%	75%	max
	radius_mean	569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000
	texture_mean	569.0	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.28000
	perimeter_mean	569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000
	area_mean	569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000
	smoothness_mean	569.0	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.16340
	compactness_mean	569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540
	concavity_mean	569.0	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.42680
	concave points_mean	569.0	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.20120
	symmetry_mean	569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400
	fractal_dimension_mean	569.0	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.09744
	radius_se	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300
	texture_se	569.0	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.88500
	perimeter_se	569.0	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.98000
	area_se	569.0	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.20000
	smoothness_se	569.0	0.007041	0.003003	0.001713	0.005169	0.006380	0.008146	0.03113
	compactness_se	569.0	0.025478	0.017908	0.002252	0.013080	0.020450	0.032450	0.13540

	radius_se	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300
	texture_se	569.0	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.88500
	perimeter_se	569.0	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.98000
	area_se	569.0	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.20000
	smoothness_se	569.0	0.007041	0.003003	0.001713	0.005169	0.006380	0.008146	0.03113
	compactness_se	569.0	0.025478	0.017908	0.002252	0.013080	0.020450	0.032450	0.13540
	concavity_se	569.0	0.031894	0.030186	0.000000	0.015090	0.025890	0.042050	0.39600
	concave points_se	569.0	0.011796	0.006170	0.000000	0.007638	0.010930	0.014710	0.05279
	symmetry_se	569.0	0.020542	0.008266	0.007882	0.015160	0.018730	0.023480	0.07895
	fractal_dimension_se	569.0	0.003795	0.002646	0.000895	0.002248	0.003187	0.004558	0.02984
	radius_worst	569.0	16.269190	4.833242	7.930000	13.010000	14.970000	18.790000	36.04000
	texture_worst	569.0	25.677223	6.146258	12.020000	21.080000	25.410000	29.720000	49.54000
	perimeter_worst	569.0	107.261213	33.602542	50.410000	84.110000	97.660000	125.400000	251.20000
	area_worst	569.0	880.583128	569.356993	185.200000	515.300000	686.500000	1084.000000	4254.00000
	smoothness_worst	569.0	0.132369	0.022832	0.071170	0.116600	0.131300	0.146000	0.22260
	compactness_worst	569.0	0.254265	0.157336	0.027290	0.147200	0.211900	0.339100	1.05800
	concavity_worst	569.0	0.272188	0.208624	0.000000	0.114500	0.226700	0.382900	1.25200
	concave points_worst	569.0	0.114606	0.065732	0.000000	0.064930	0.099930	0.161400	0.29100
	symmetry_worst	569.0	0.290076	0.061867	0.156500	0.250400	0.282200	0.317900	0.66380
	fractal_dimension_worst	569.0	0.083946	0.018061	0.055040	0.071460	0.080040	0.092080	0.20750

Each column of the output means:

1. count: This column tells you the number of non-null values in each numerical column.

It provides a count of available data points for each column.

2. mean: This column shows the mean (average) value for each column. For example, in the "radius_mean" column, it indicates the average radius of the tumor.

3. std: The standard deviation measures the variability or dispersion of data. It shows how much the values in each column deviate from the mean. A higher standard deviation indicates more variation.

4. min: This column displays the minimum value in each column. For example, in the "radius_mean" column, it indicates the lowest radius of the tumor among all the tumor radii.

5. 25%: This row represents the 25th percentile value. It shows the value below which 25% of the data falls. It's often referred to as the first quartile (Q1).

6. 50%: This row corresponds to the median value (50th percentile) in each column.

The median is the middle value when the data is sorted.

7. 75%: This row shows the 75th percentile value, which is also known as the third quartile (Q3). It represents the value below which 75% of the data falls.

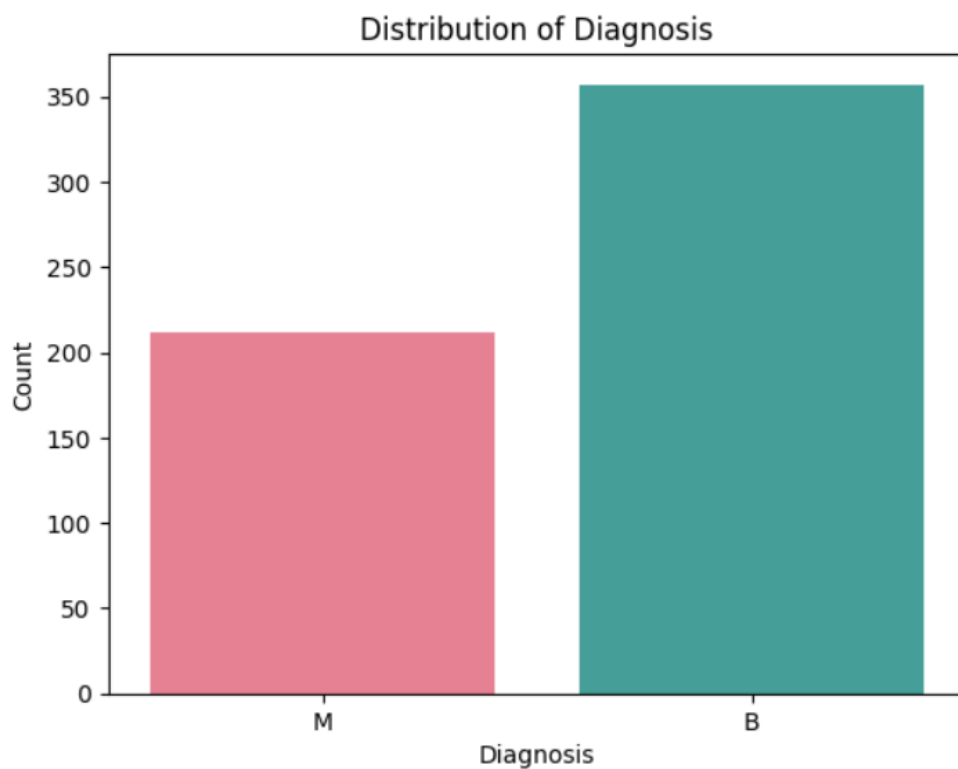
8. max: This row displays the maximum value in each column. For example, in the "radius_mean" column, it indicates the highest radius of the tumor.

An overview of the central tendency, spread, and the range of values for the numerical columns in the mobile phone dataset is obtained to understand key statistics about the data.

Analysis:

Calculating how many patients have malignant and benign tumors from the given dataset.

```
[5] import seaborn as sns  
df["diagnosis"].value_counts()  
  
B    357  
M    212  
Name: diagnosis, dtype: int64
```

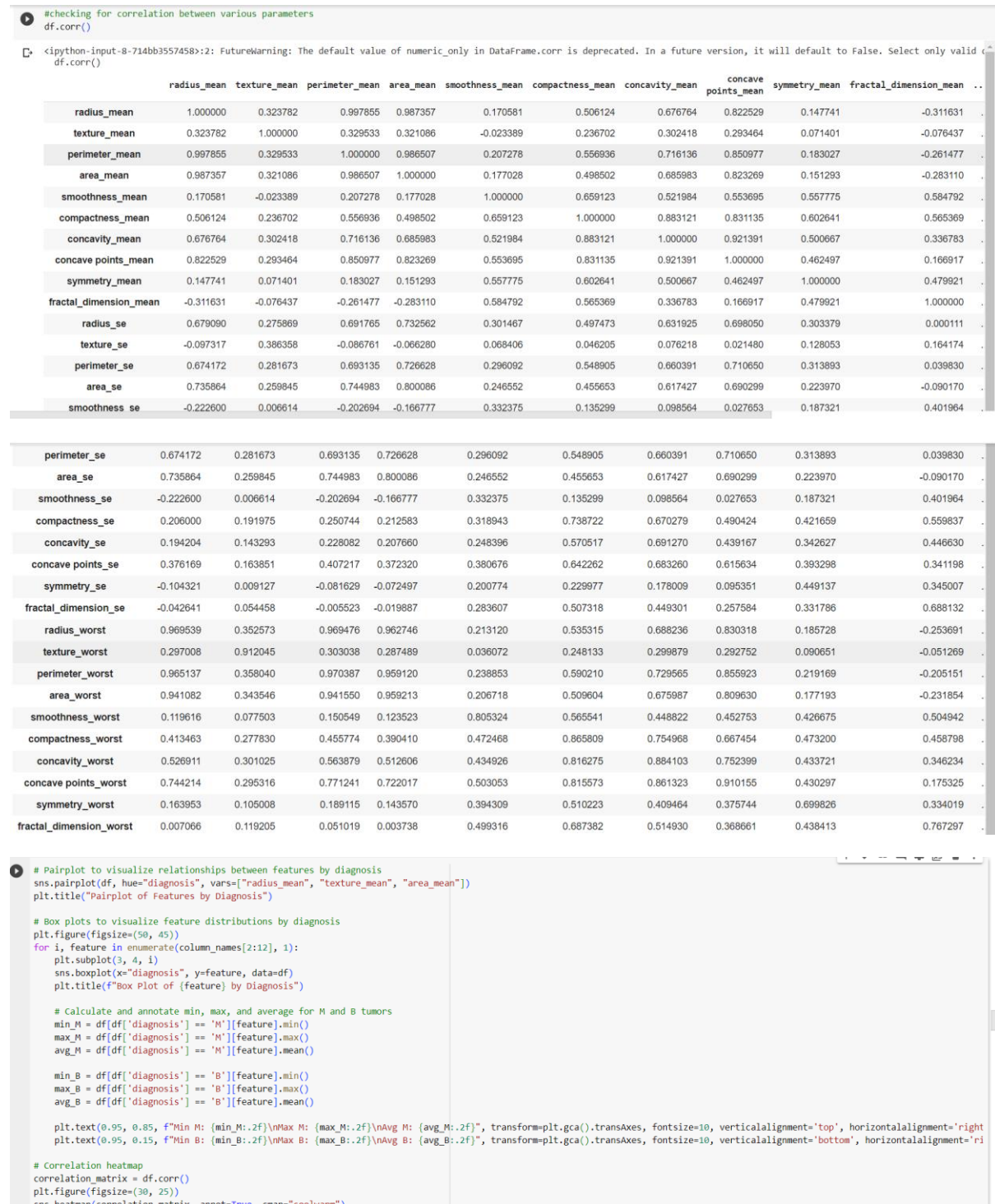


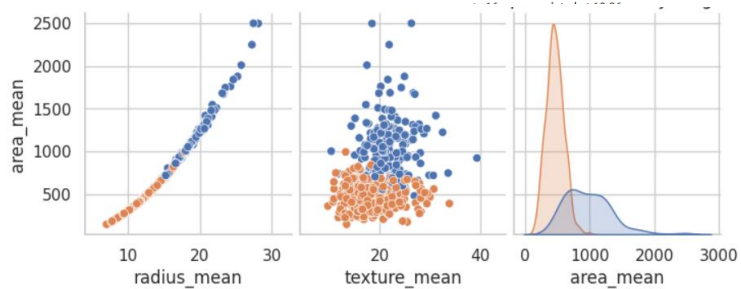
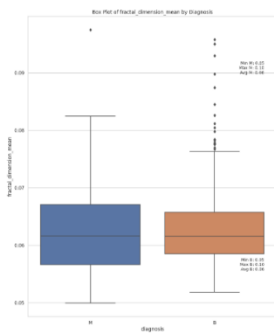
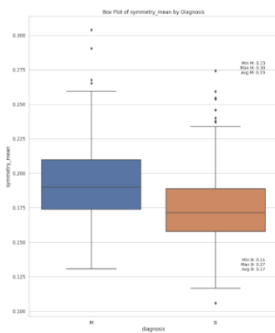
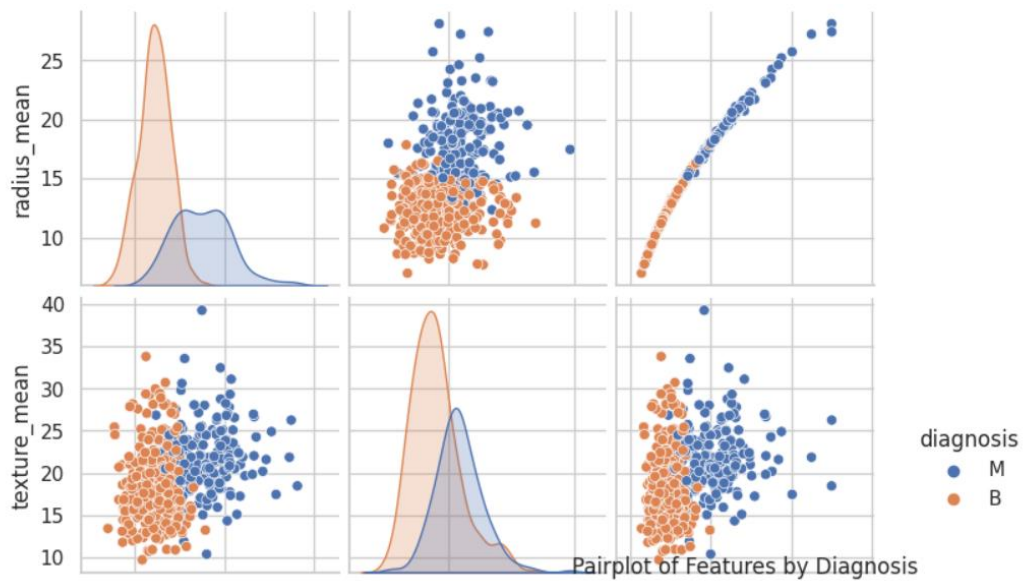
So,

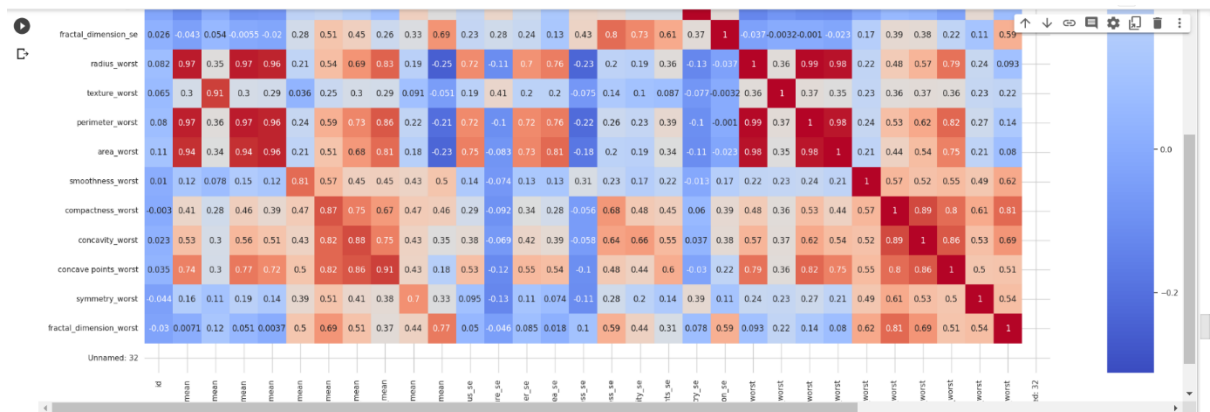
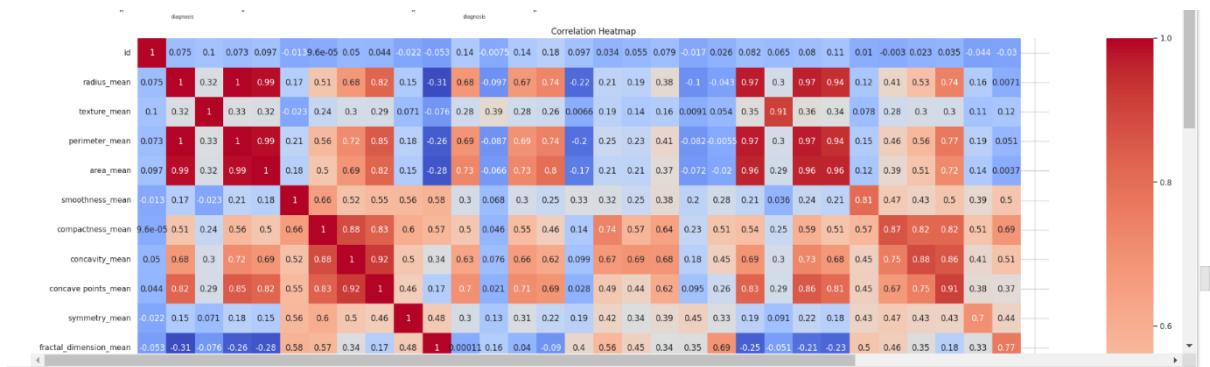
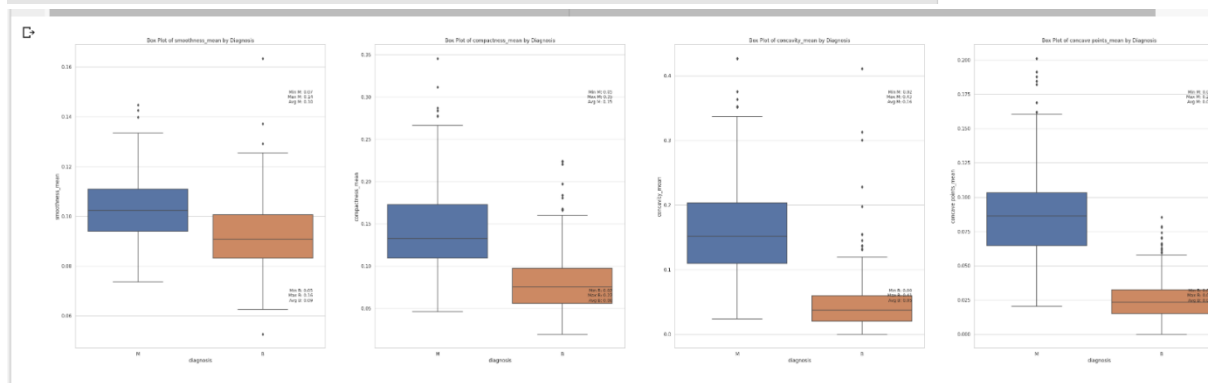
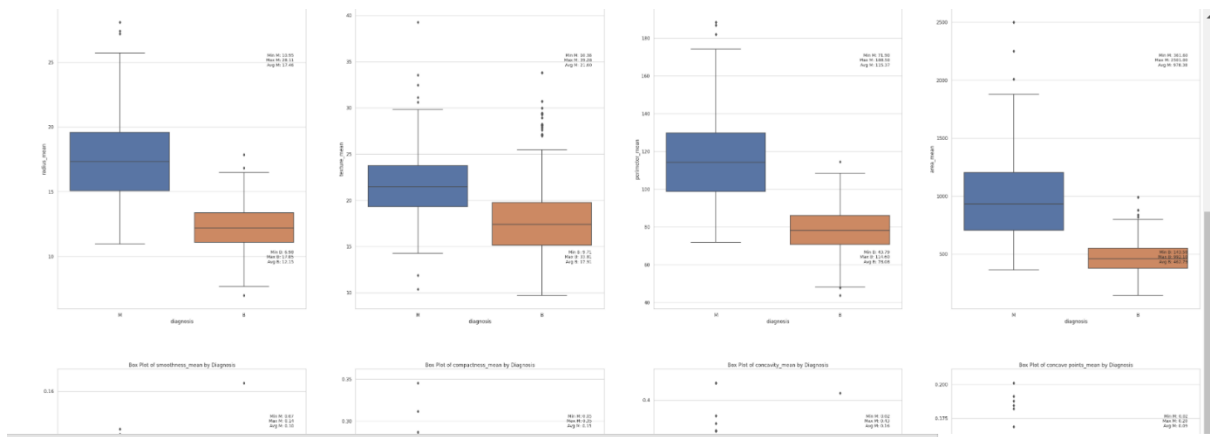
B=357

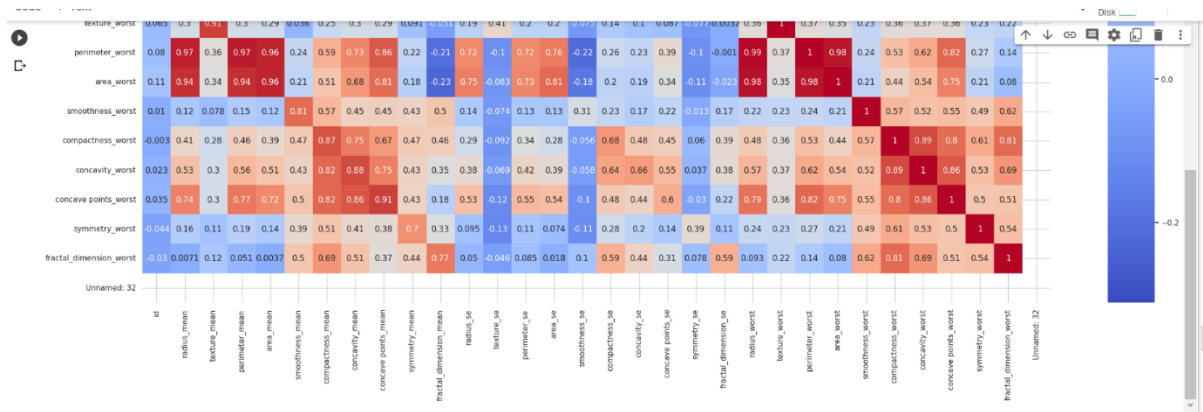
M=212

We will now look at the features of each and the correlation with diagnosis which might help us detect the type of cancer and streamline their treatment.









Above are various graphs and heat maps that show relation of diagnosis with all the other features.

1. Positive Correlation: A positive correlation coefficient (close to 1) between two features indicates that as one feature increases, the other tends to increase as well. In the context of breast cancer diagnosis, this might imply that two features tend to increase together in cancerous or benign tumors.

2. Negative Correlation: A negative correlation coefficient (close to -1) between two features indicates that as one feature increases, the other tends to decrease, and vice versa. In the context of breast cancer diagnosis, this might suggest an inverse relationship between two features.

3. No Correlation: A correlation coefficient close to 0 indicates little to no linear relationship between two features. In other words, changes in one feature do not have a strong influence on the other.

By examining the correlation matrix, we can identify which features are strongly correlated with each other. This information can be valuable for feature selection for gaining insights into the relationships between different attributes in the dataset. It can also help you identify potential multicollinearity issues if multiple features are highly correlated.

In the heat map above we can identify the correlations of the features with each other and decide how to adjust the treatments.

An example of correlation with radius_mean-

Correlation with Radius_Mean (in descending order):

radius_mean	1.000000
perimeter_mean	0.997855
area_mean	0.987357
radius_worst	0.969539
perimeter_worst	0.965137
area_worst	0.941082
concave_points_mean	0.822529
concave_points_worst	0.744214
area_se	0.735864
radius_se	0.679090
concavity_mean	0.676764
perimeter_se	0.674172
concavity_worst	0.526911
compactness_mean	0.506124
compactness_worst	0.413463
concave_points_se	0.376169
texture_mean	0.323782
texture_worst	0.297008
compactness_se	0.206000
concavity_se	0.194204
smoothness_mean	0.170581
symmetry_worst	0.163953
symmetry_mean	0.147741
smoothness_worst	0.119616
id	0.074626
fractal_dimension_worst	0.007066
fractal_dimension_se	-0.042641
texture_se	-0.097317
symmetry_se	-0.104321
smoothness_se	-0.222600
fractal_dimension_mean	-0.311631

In this we can see high positive relation with perimeter, low positive with symmetry and negative with smoothness.

Mean by diagnosis-

This metric can help identify tumor type of future patients in its early stages.

```
Means by Diagnosis:
,      radius_mean  texture_mean  perimeter_mean  area_mean  \
diagnosis
B      12.146524    17.914762      78.075406   462.790196
M      17.462830    21.604906     115.365377   978.376415

      smoothness_mean  compactness_mean  concavity_mean  \
diagnosis
B           0.092478          0.080085          0.046058
M           0.102898          0.145188          0.160775

      concave points_mean  symmetry_mean  fractal_dimension_mean  ...  \
diagnosis
B           0.025717          0.174186          0.062867  ...
M           0.087990          0.192909          0.062680  ...

      radius_worst  texture_worst  perimeter_worst  area_worst  \
diagnosis
B      13.379801     23.515070      87.005938   558.899440
M      21.134811     29.318208     141.370330  1422.286321

      smoothness_worst  compactness_worst  concavity_worst  \
diagnosis
B           0.124959          0.182673          0.166238
M           0.144845          0.374824          0.450606

      concave points_worst  symmetry_worst  fractal_dimension_worst
diagnosis
B           0.074444          0.270246          0.079442
M           0.182237          0.323468          0.091530

[2 rows x 30 columns]
```

In the above snapshot we can see mean of all features by diagnosis.

Risk Score-

We can calculate the risk score of each patient and prioritize accordingly as the relative number of doctors to patients is low.

```

# Calculating the risk score for each row (patient) in the DataFrame
def calculate_risk(row):
    risk = 0
    for feature, weight in factors.items():
        risk += row[feature] * weight
    return risk

# Applying the calculate_risk function to each row
df['risk_score'] = df.apply(calculate_risk, axis=1)

df_sorted = df.sort_values(by='risk_score', ascending=False)

# Printing the DataFrame with risk scores
print(df_sorted[['diagnosis', 'risk_score']])
|

```

```

┌┐  diagnosis  risk_score
461      M      8.116420
180      M      7.636470
82       M      7.540315
212      M      7.474710
239      M      7.424906
...      ...
504      B      3.148770
166      B      3.135797
525      B      3.029380
59       B      2.907476
101      B      2.745050

```

Here, we use only radius_mean, texture_mean and smoothness_mean to keep the analysis concise. The relative weights of these three parameters are taken from National Library of Medicine US.

Managerial Implications-

1. **Early Detection Matters:** The analysis may reveal that certain features are highly indicative of malignancy. Understanding which features are critical can underscore the importance of early detection and regular screenings in healthcare management.
2. **Resource Allocation:** The analysis demonstrates that we can utilize available resources to appropriately prioritize high risk patients but also looking after low risk patients and keep a constant watch in case of an unexpected occurrence.
3. **Risk Assessment and Insurance:** Insights from the analysis can be applied to risk assessment models in insurance or healthcare finance. We gain an understanding of how data-driven risk assessment can help insurers set premiums and coverage policies more accurately.
4. **Medical Technology Investments:** The analysis highlights the significance of certain diagnostic features, it can encourage discussions about investing in new medical technologies or equipment.

5. Strategic Partnerships: Insights from the analysis might suggest opportunities for strategic partnerships between healthcare providers, diagnostic labs, and technology companies for smooth and efficient testing and delivery ensuring timely diagnosis that can save more lives.

Overall, the dataset provides insights into how various features are correlated and how by studying them we can detect whether a tumor is malignant or benign in very early stages which increases the chances of maintaining the patient's health and also increases the trust in medical institutes which in turn increases the brand equity of the health industry which will be beneficial the medical institutes as well.