Wisconsin Breast Cancer Dataset

by Rajshri Ganesh Iyer

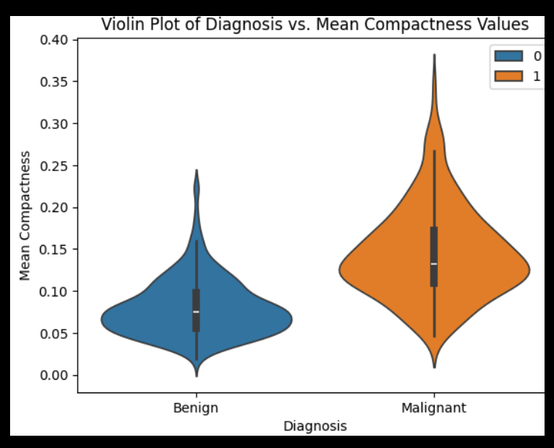
Dataset from : UCI Machine Learning Repo

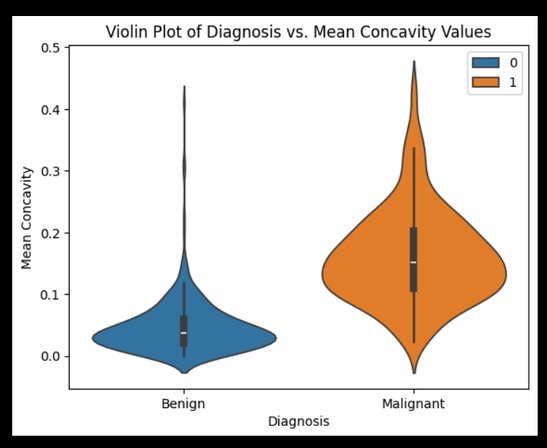
Link to Kaggle Dataset:

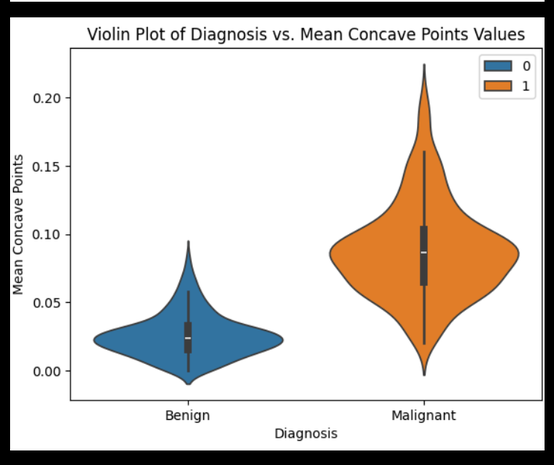
https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

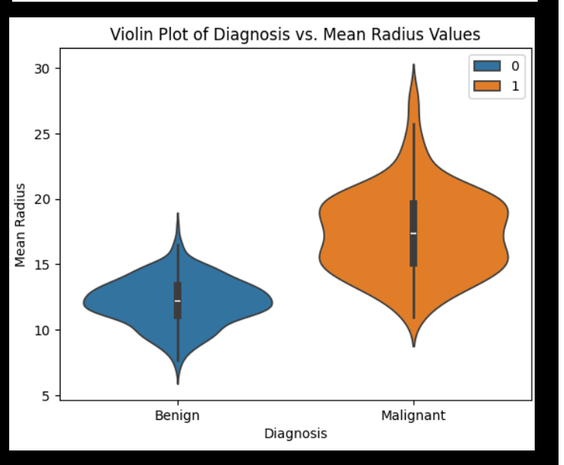
OBJECTIVE: To identify the features that have highest correlation with breast cancer diagnosis and to fit a Linear Regression model to the data for breast cancer diagnosis prediction.

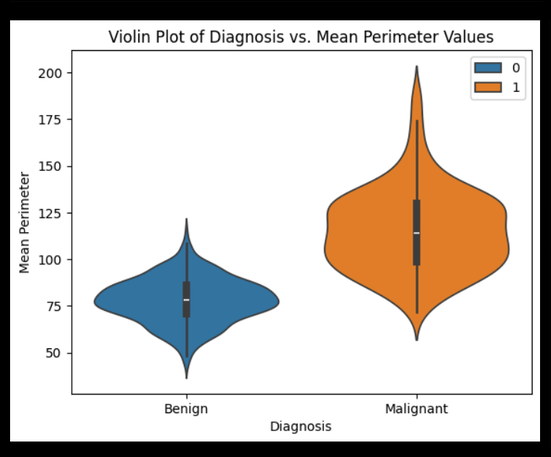
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	0.062798
std	1.250206e+08	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	0.007060
min	8.670000e+03	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.049960
25%	8.692180e+05	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	0.057700
50%	9.060240e+05	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	0.061540
75%	8.813129e+06	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	0.066120
max	9.113205e+08	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	0.097440

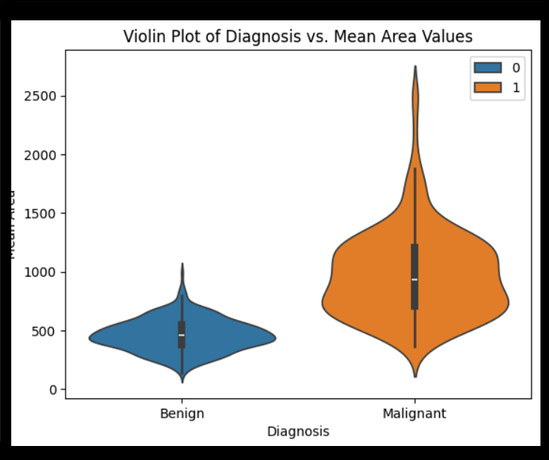


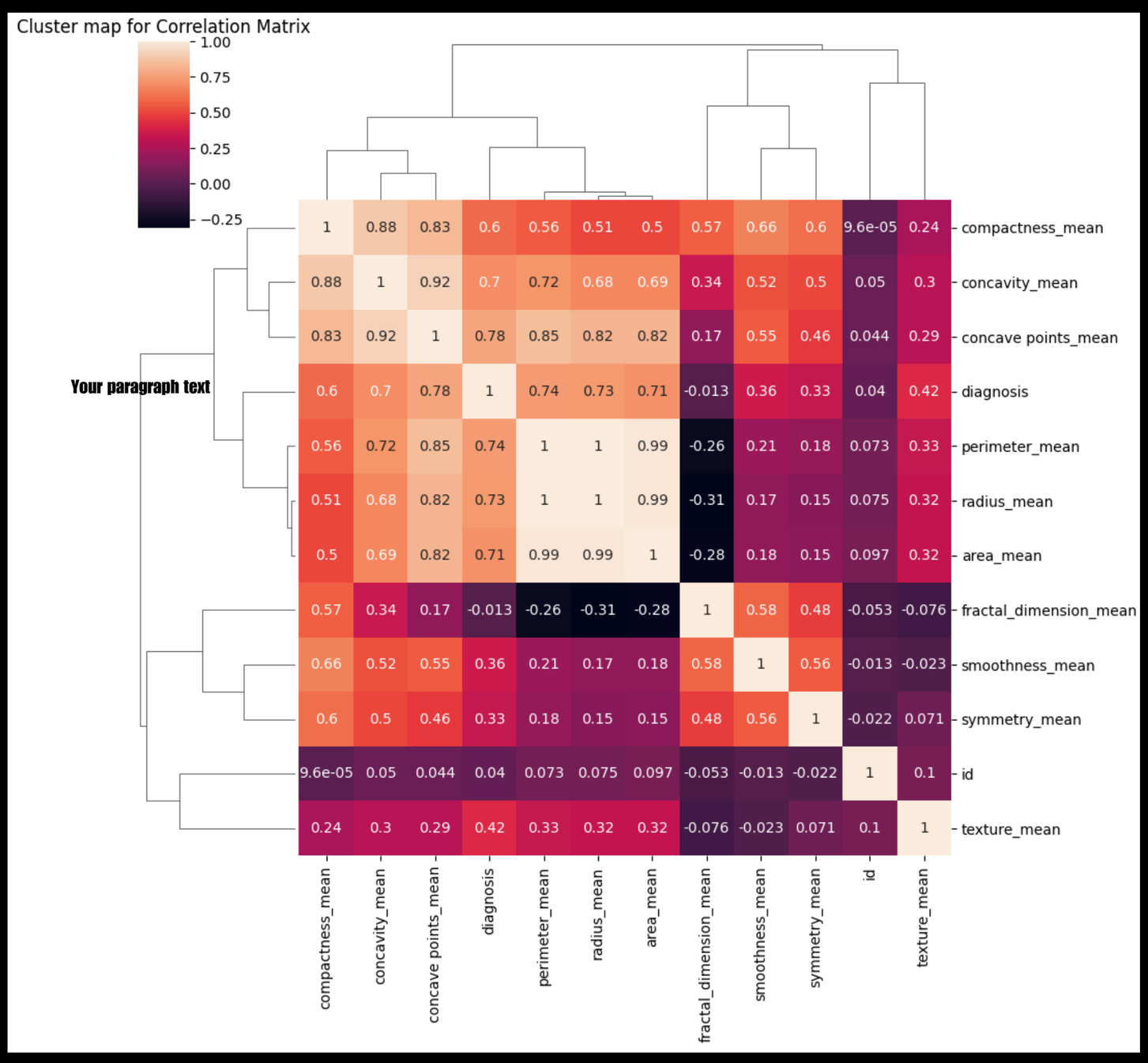




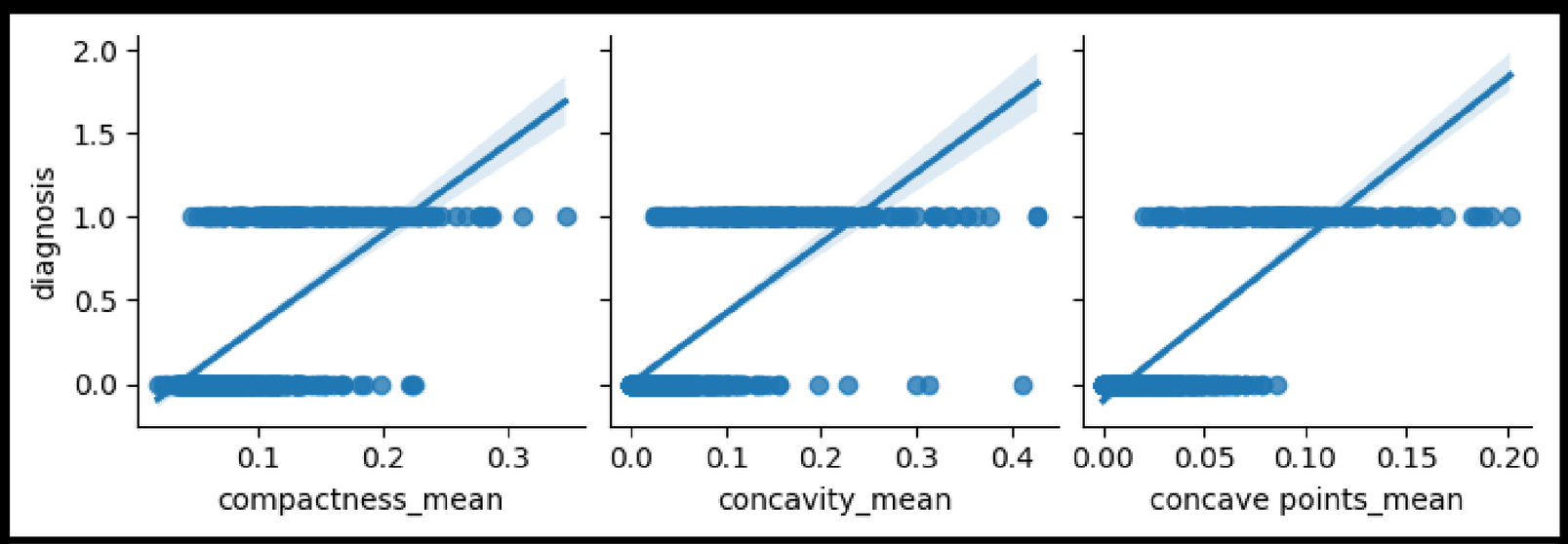






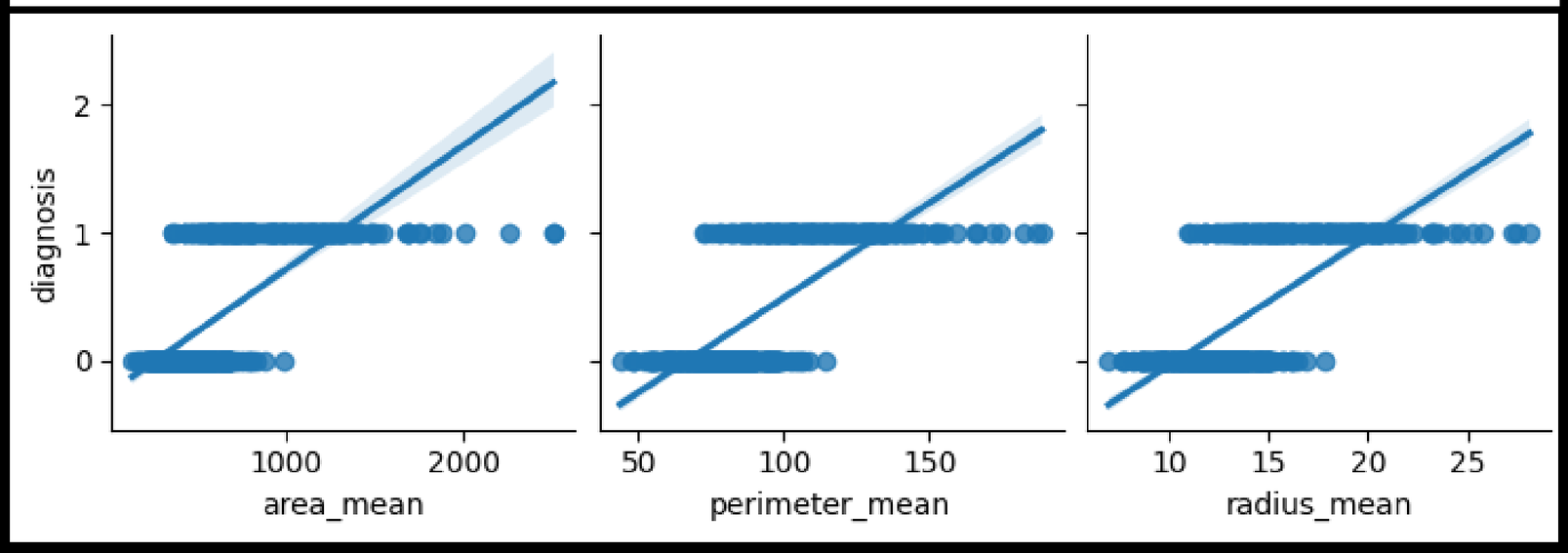


Your paragraph text



Steps

1. Extract the features to be used for fitting
2. Split data into training and testing
3. Fit a Linear Regression model
4. Extract the model parameters for R2, MSE, intercept and coefficients & evaluate the model



5. Plot the Linear Regression data

MODEL PARAMETERS

```
Mean squared error: 0.08494739844208647
R-squared: 0.6383962036838009
b0 (model intercept)          = -1.3346645166241813
b1 (coeff for compactness)    = 0.4225071269807399
b2 (coeff for concavity)      = 0.6485696045302635
b3 (coeff for concave points)= 8.470875669627944
b4 (coeff for perimeter)      = -0.05294643956330038
b5 (coeff for radius)         = 0.4769109453062004
b6 (coeff for area)           = -0.0010274395335266862
```

CONCLUSION: Based on Correlation Matrix, only 6 out of 10 features have high correlation with breast cancer diagnosis. Linear Regression model has a low R2 value, hence a better model needs to be used for fitting.