Data Visualization of Breast Cancer Data

Francis Osei

17/12/2021

In this project we are going to explore the breast cancer data from the UCI Machine learning repository (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).

1. Outline of the Project

The figure below shows the the overall visualization process of our data.

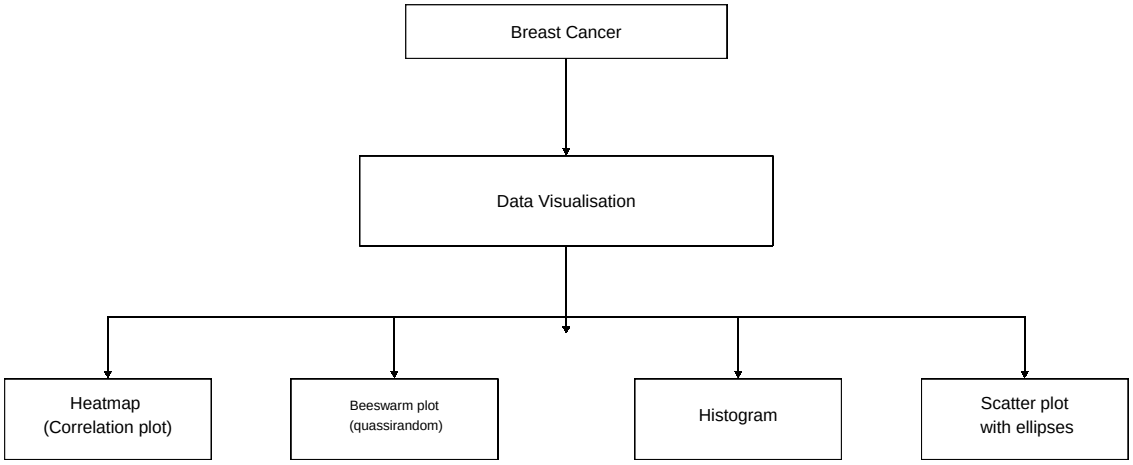

figure 1: Overall process flow of breast cancer data Visualisation

Source: Peter Higgins
1/26/2019

2. Exploring the features of Breast Cancer Data

Figure 2 shows all the features (Attributes) of our data. Each feature has either mean, Standard error (se) and worst. This gives us a clear idea of the features we are working with.

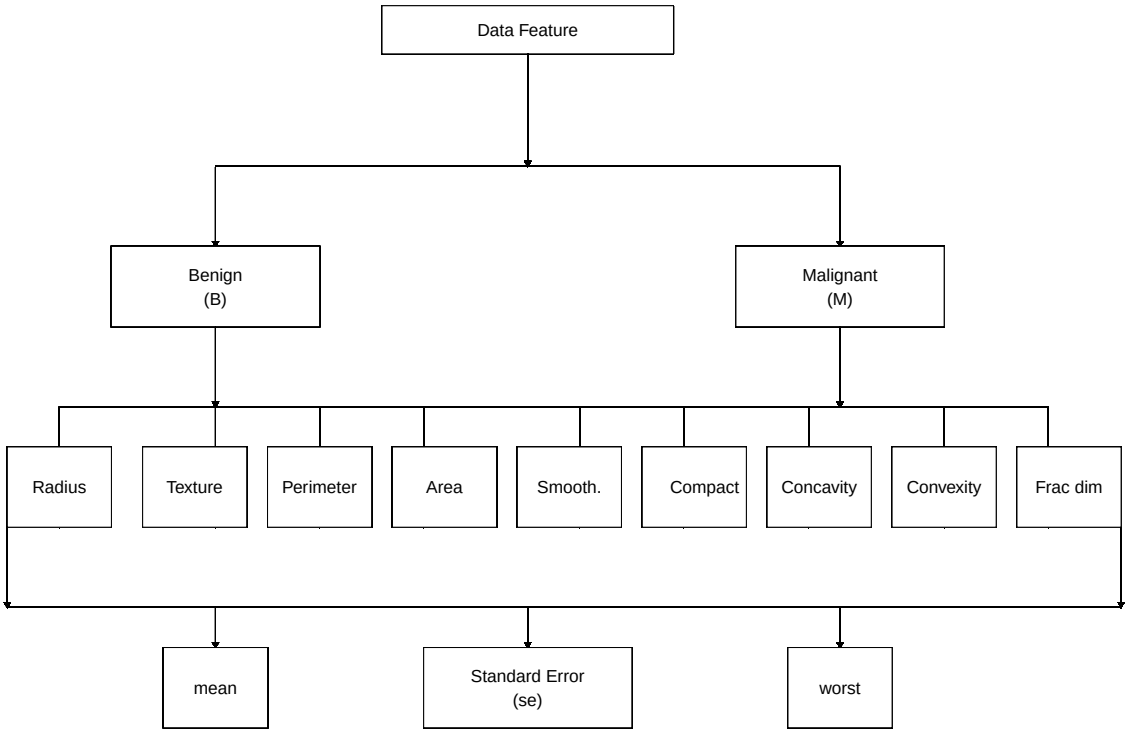

figure 2: Data representation

3. Data Visualization

In this section we will provide a graphical representation of our data. This part provides an accessible way to understand the patterns,outliers and trends in breast cancer data.

3.1. Heatmap (Correlation Plot)

The correlation plot uses the pearson correlation to show the direction and strength of the relation of our features. This helps to reduce the number of features in our data. Here we represented the correlation in form of heatmap showing the range of values using colours. This visualization is important especially when doing regression analysis.

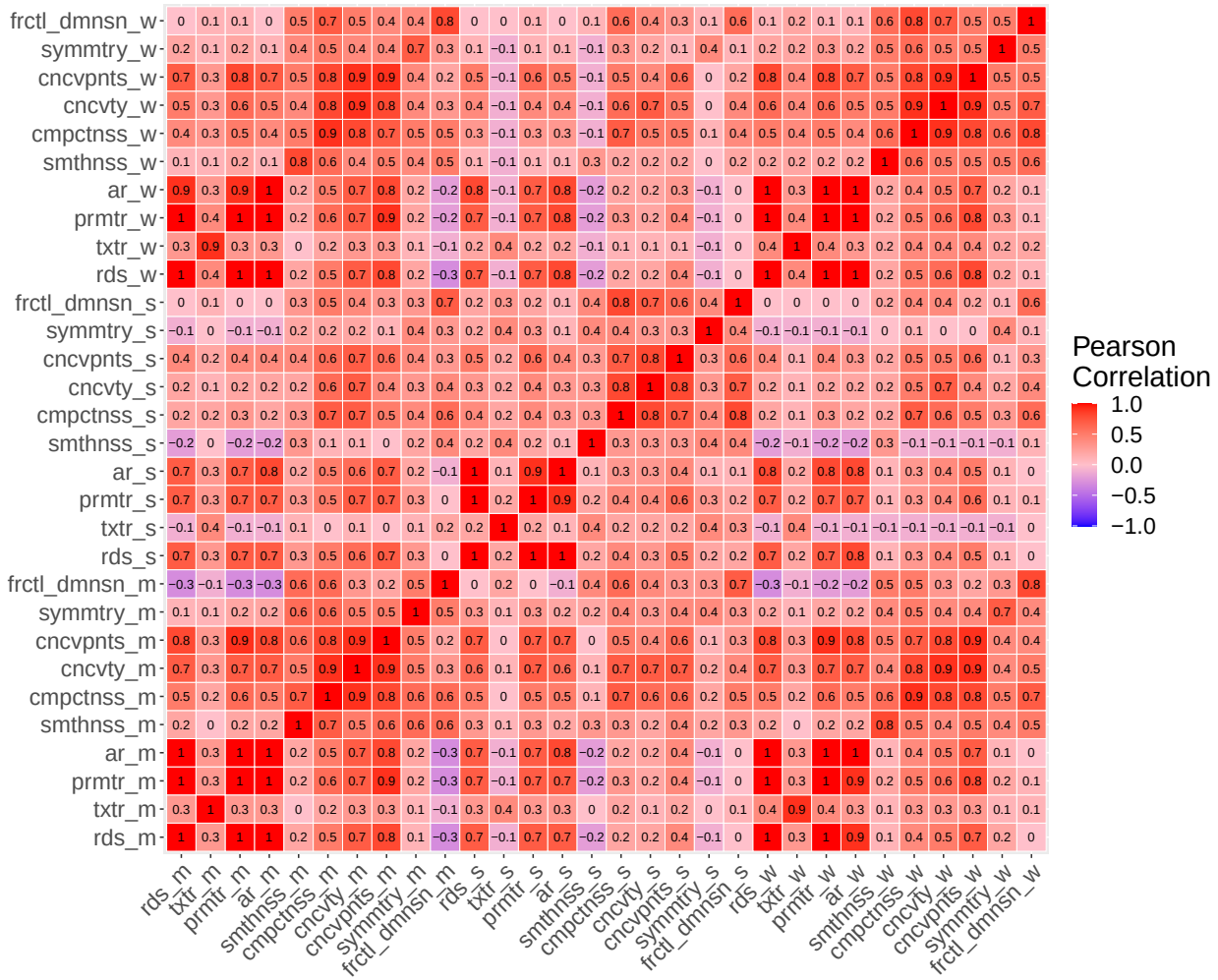

figure 3: Heatmap (correlation plot)

3.2. The Beeswarm plot (Quasirandom)

The beeswarm is a techique to avoid overplotting. This randomly move data points away from each other. We plotted a quasirandom beeswarm to distinguish between Benign and Malignant. For each plot we have 10 features. The are some features that are difficult to differentiate between Benign and Malignant. This techniques provides a visible and easy to interpret since each data point do not overlap.

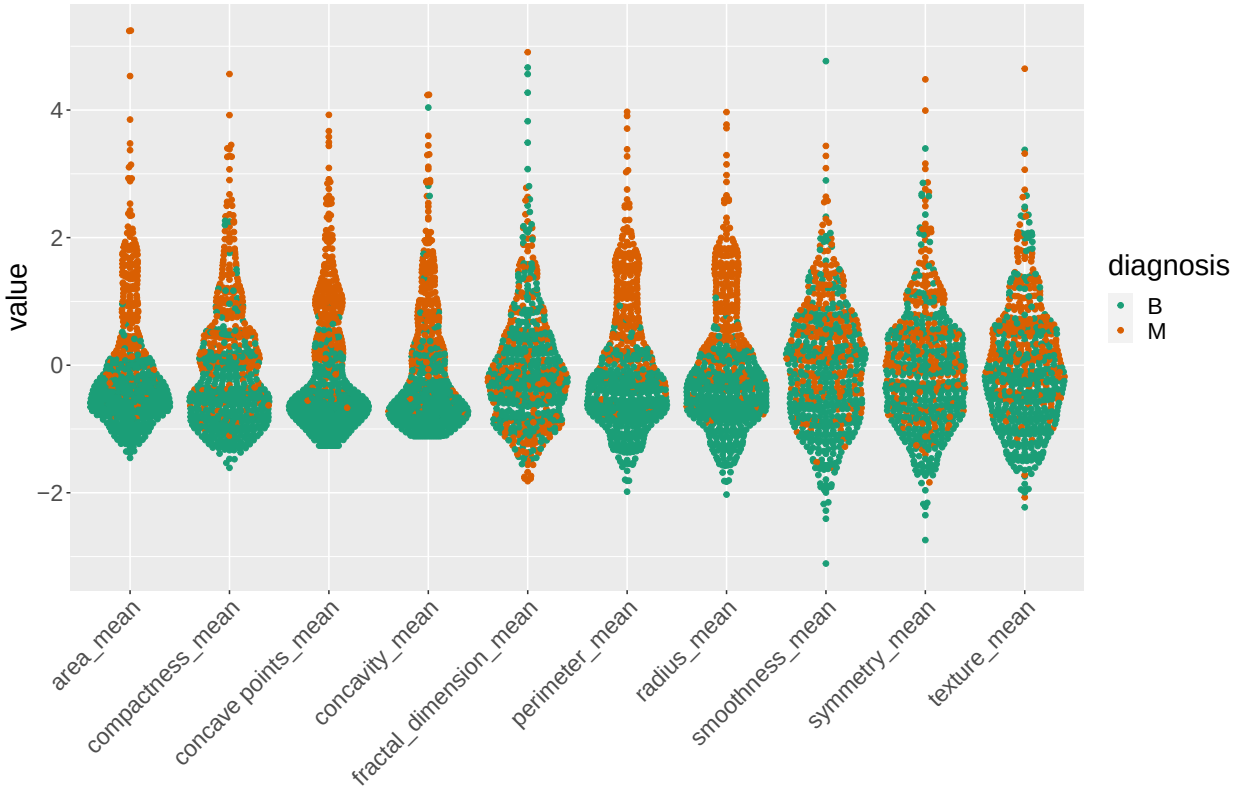

Figure 4: Beeswarm plot for first ten features

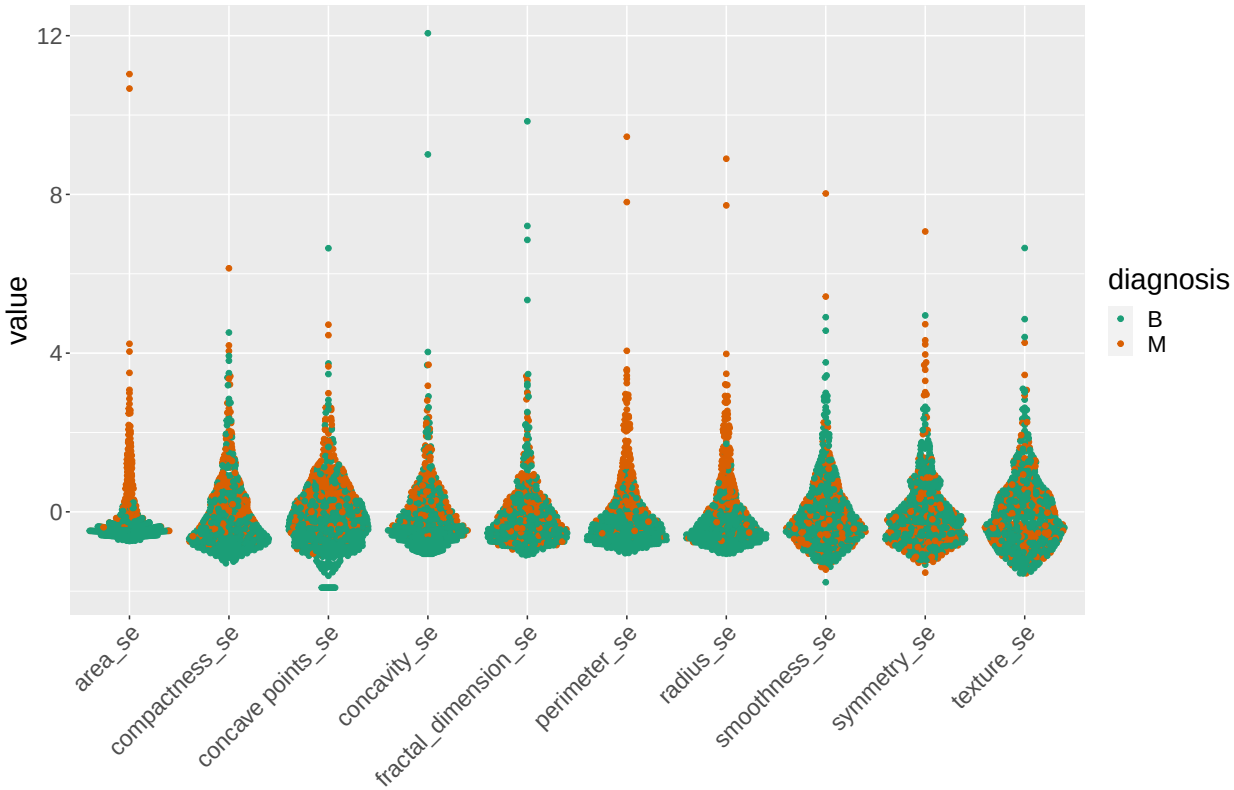

Figure 5: Beeswarm plot for middle ten features

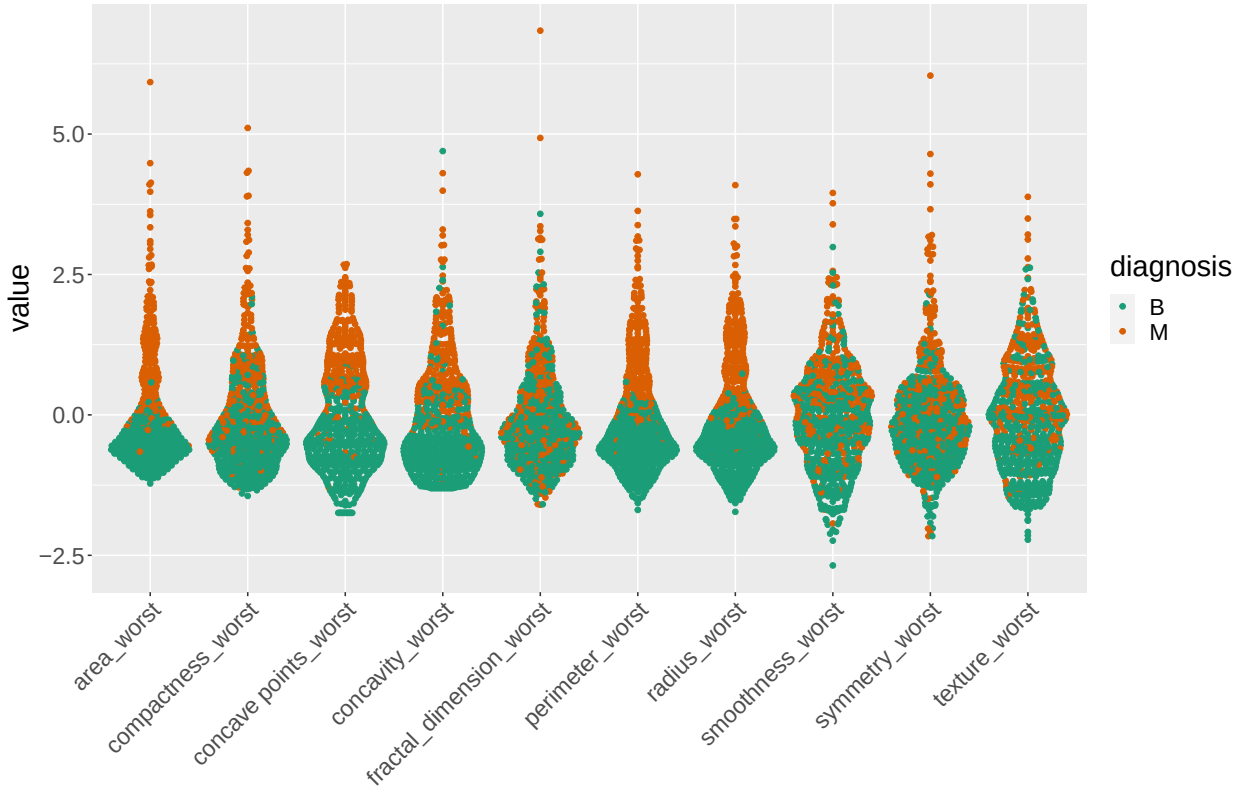

Figure 6: Beeswarm plot for last ten features

3.3. The Histogram

Some plots shows clear differentiation between the Malignant and Benign cell whiles others are very difficult to distinguish. We tried to use different plot to see if we can get a clear picture of the the difference between Benign and Malignant. In the histogram plot, the observation in each bin is represented by the height of the bar. In this plot, the variance for some features can be seen more clearly interns of Diagnosis.

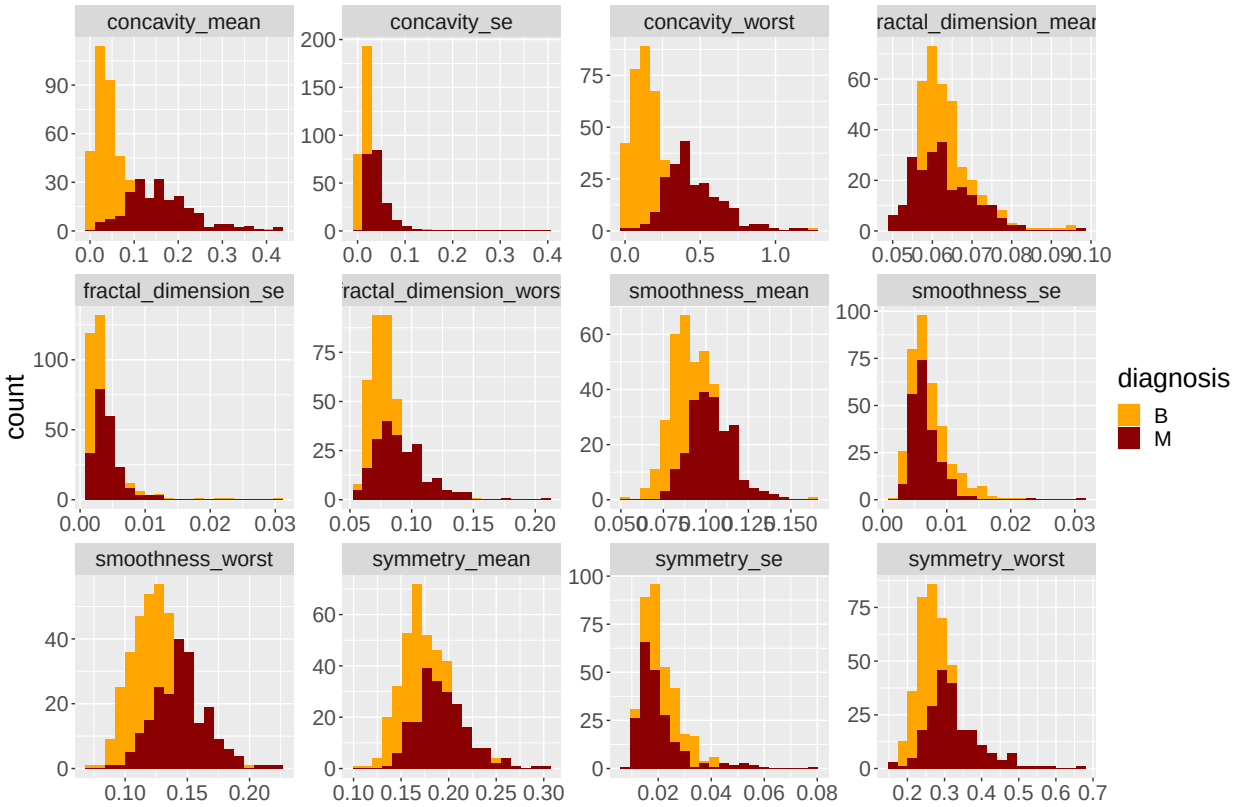

fig 7: Difficult features to Differentiate among Diagnosis

3.4. Scatter plot with individual ellipse

The scatter plot uses dots to represent two different numerical features. The scatter plot shows the relationship between the variable (mean,standard error, and worst).

For classification purpose, we define the region that contains 95% of the samples for each diagnosis. The individual ellipse provides us with some interesting difference between the diagnosis. - The two ellipse are overlapping and parallel. - The two ellipse are overlapping and perpendicular. - The two ellipse are non overlapping and parallel. - The two ellipse are non overlapping and perpendicular.

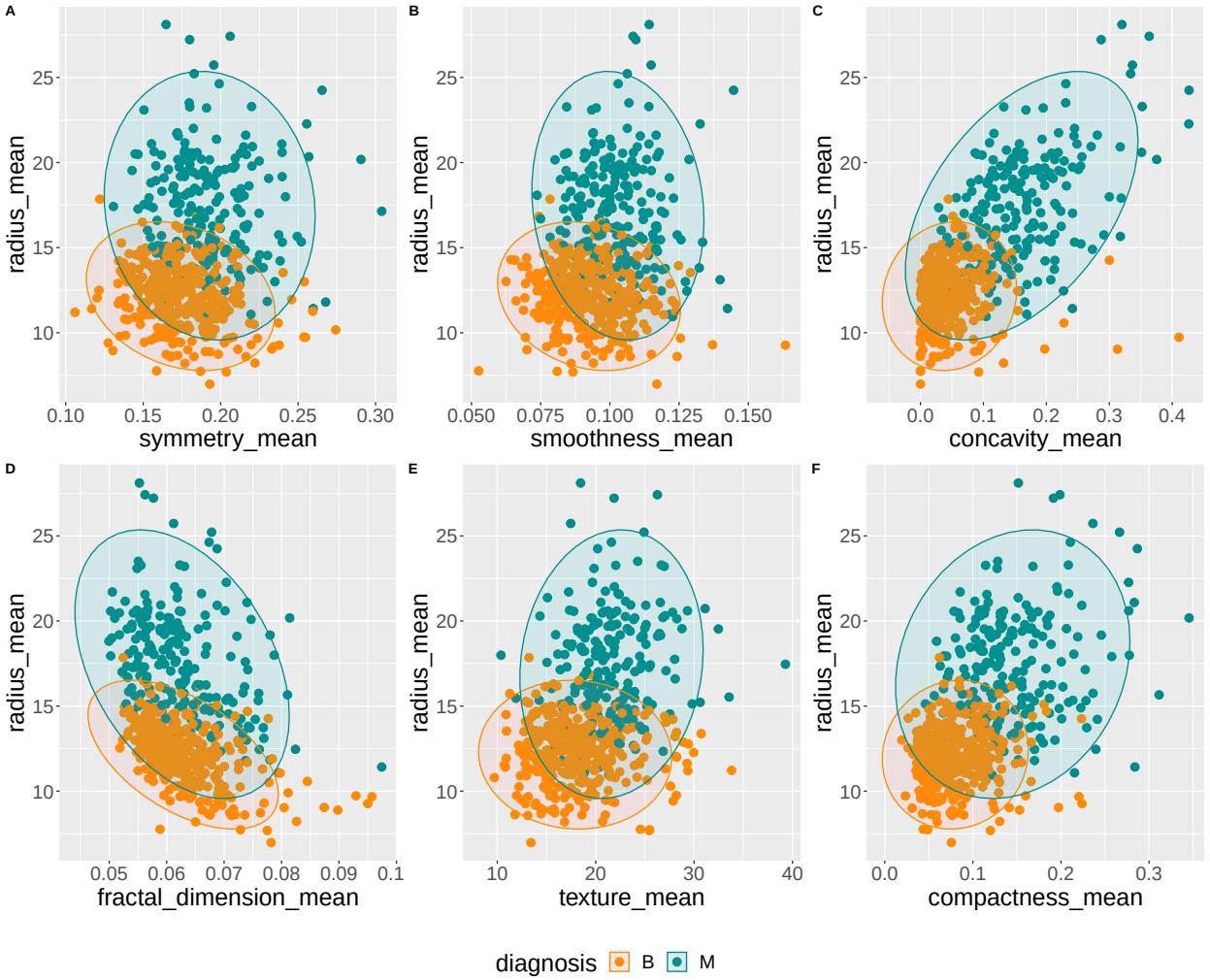

fig 8: 95% confidence region for mean features

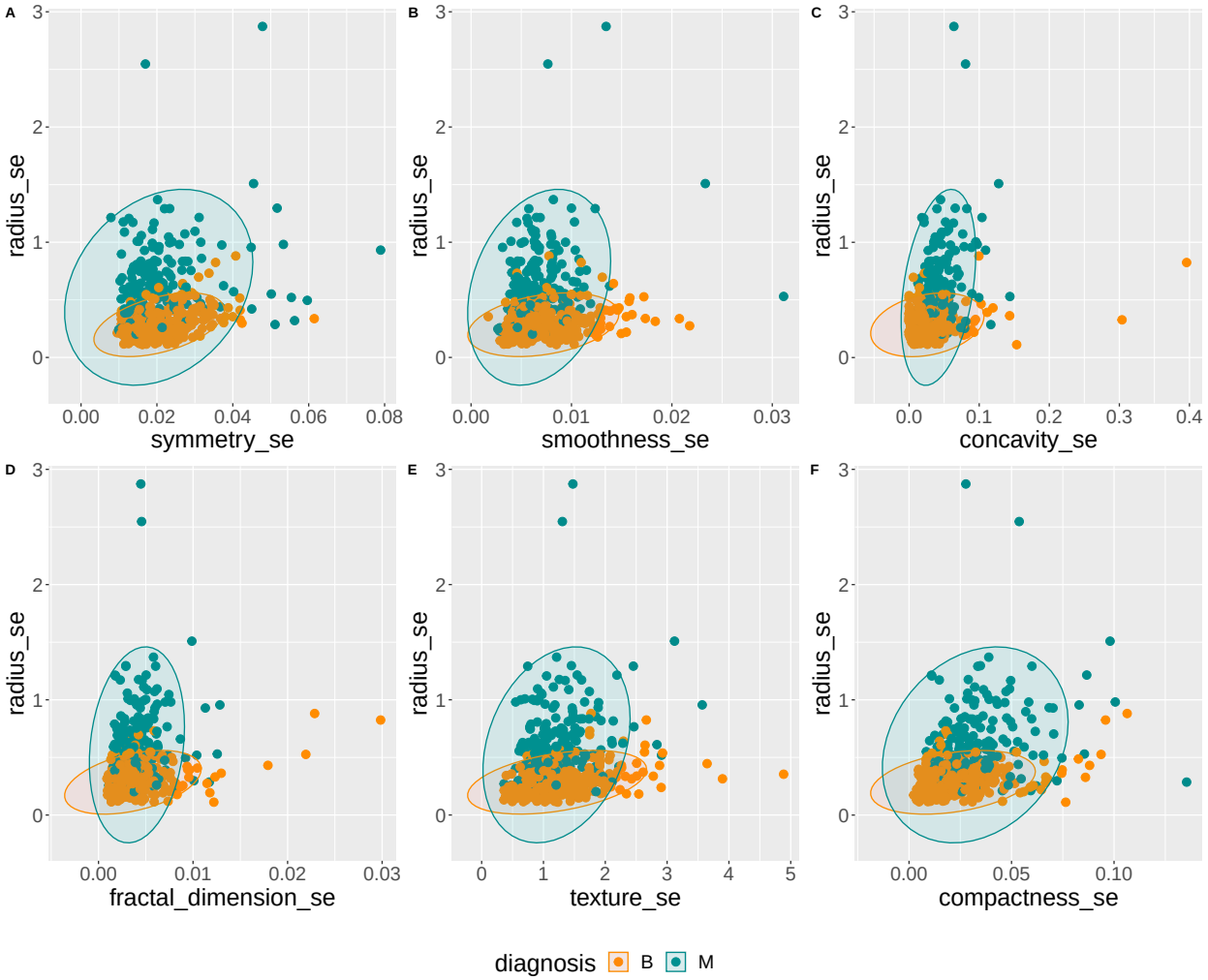

fig 9: 95% confidence region for se features

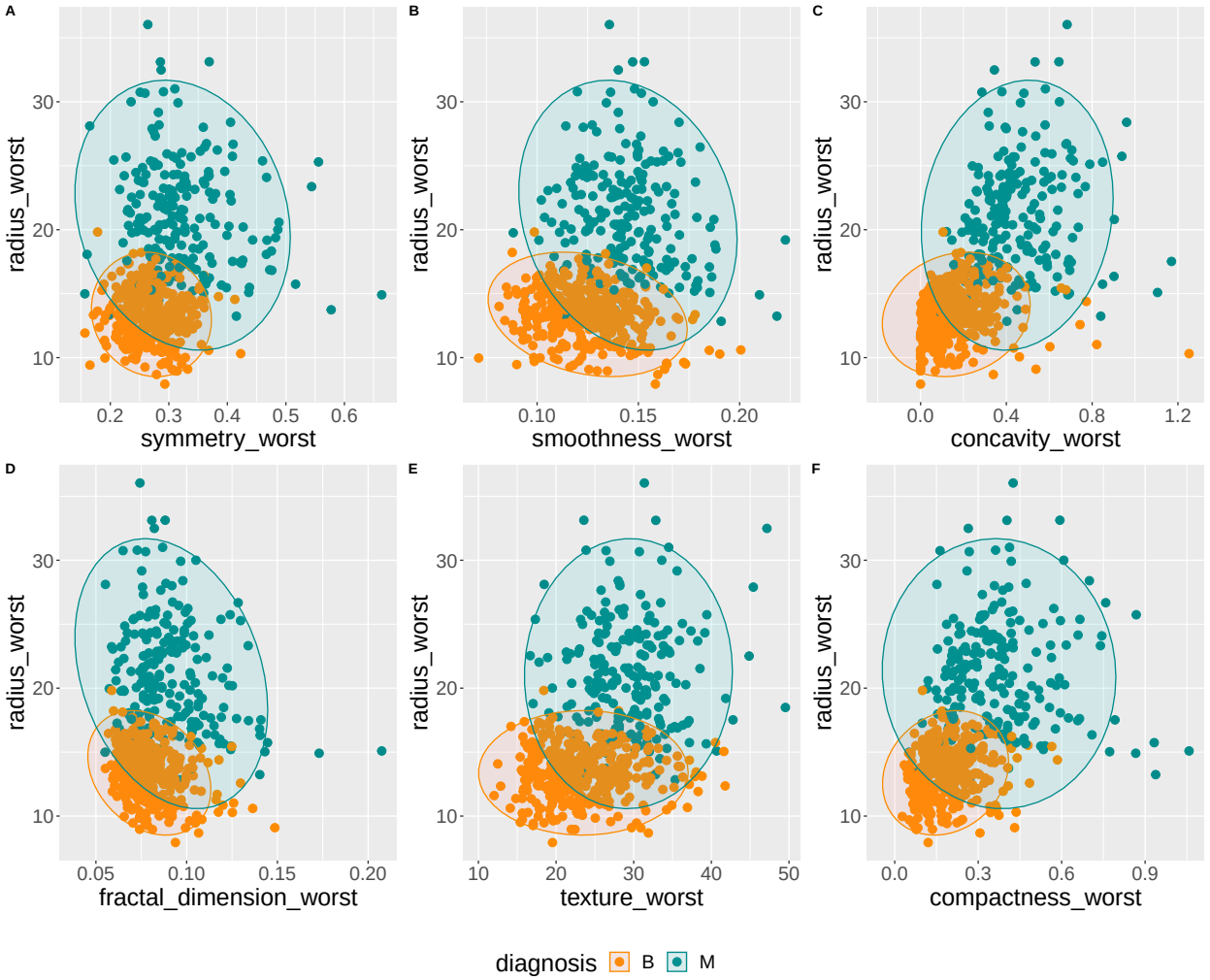

fig 10: 95% confidence region for worst features

4. Conclusion

Data visualization provides us with a clear understanding of our data. From the various plots, we noticed that Visualizing your data can help eliminate some features and also find the most useful features.