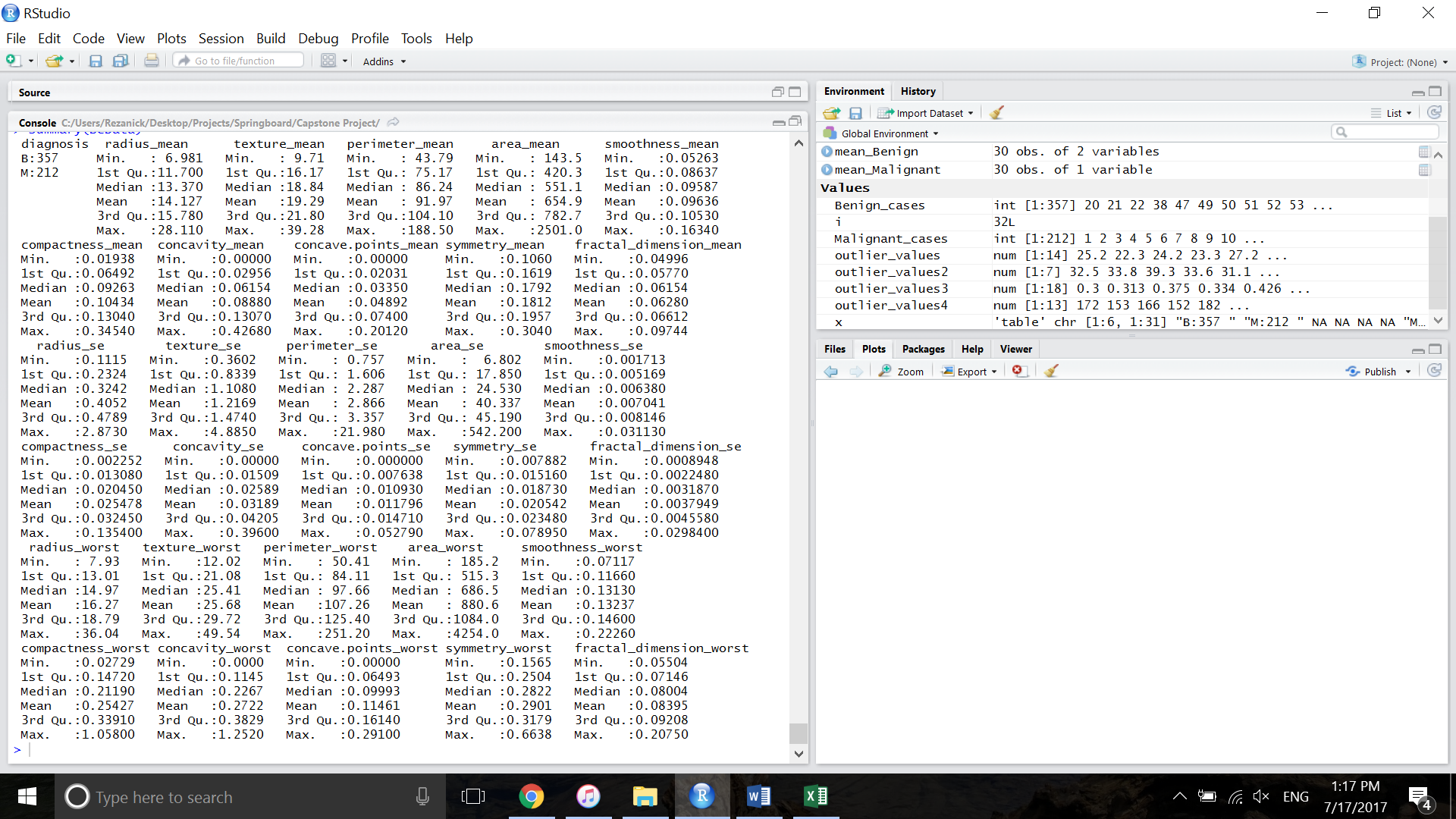
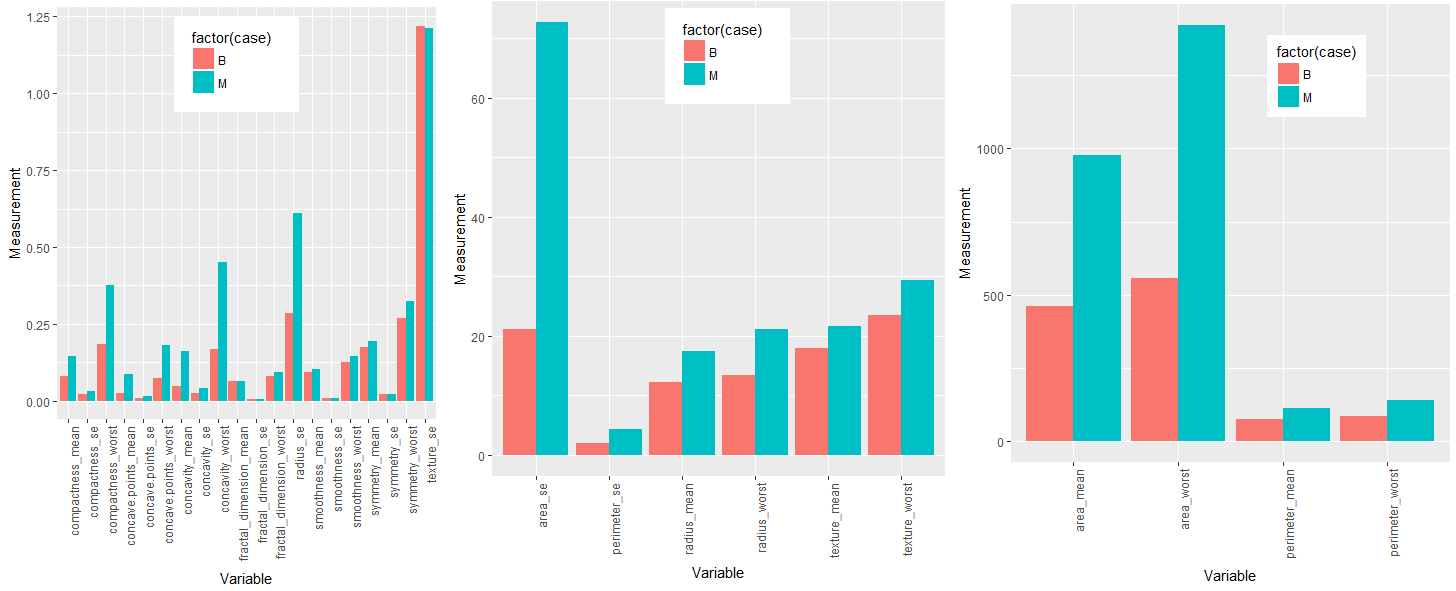
**Statistics Part:**

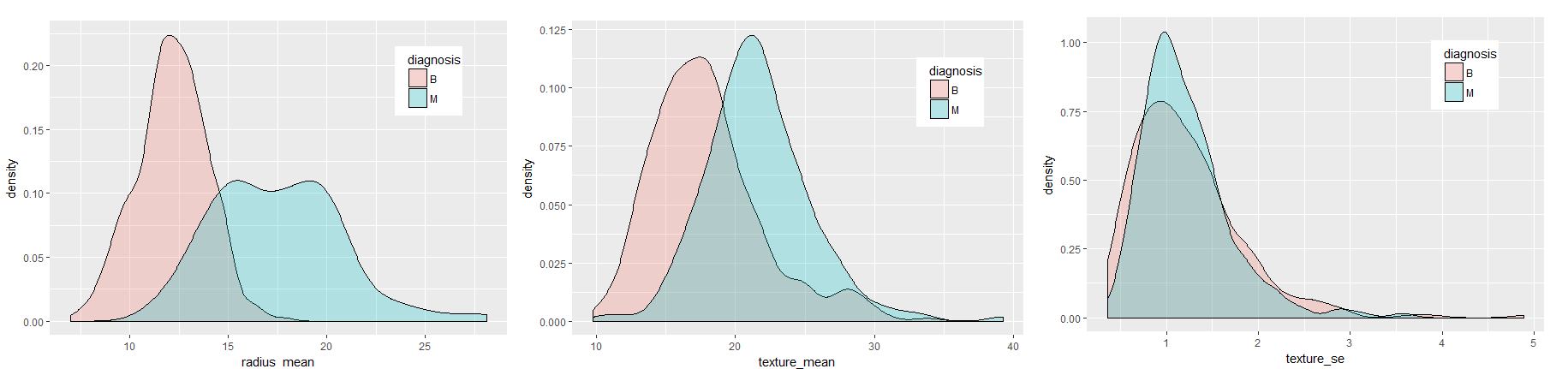
We summarize the values for each of the variables in our data set in the table below:



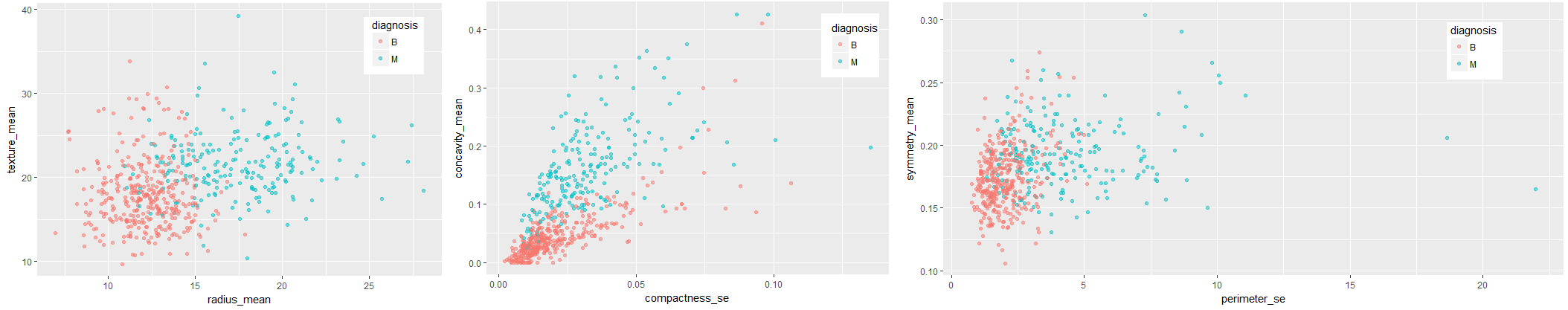
The graph below represents the mean value of the variables in B and M cases. The graphs are plotted in 3 separate graphs as the scales of variables are different. Generally speaking, variables have higher values in M cases compared to B cases (green bars vs. red bars) which suggest these computed features have potential to distinguish between B and M cases.



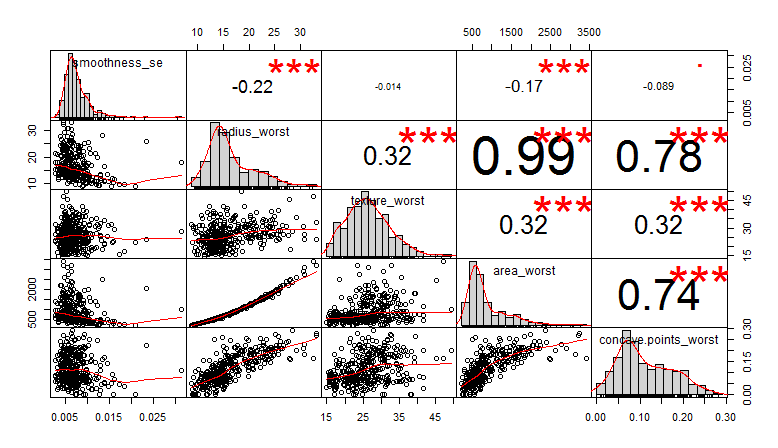
The objective of this capstone project is to develop predictive models that could distinguish between benign (B) and malignant (M) cases. Therefore, it would be interesting to plot the variables for the two groups (B and M) to identify the potential predictors for classification. Here, as an example, we show density plots of three of the variables to see their distribution in B and M cases. As seen here, radius-mean values differs between B and M cases and could be used to distinguish the groups, whereas, texture\_se value distribution is similar in both cases.

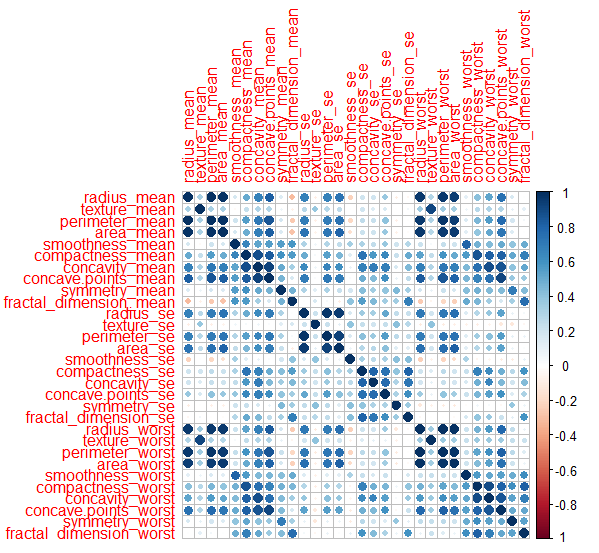


We also created scatterplots of a subset of variables to assess the distribution in B and M cases and visually assess if the variables are correlated. These scatterplots suggest that the two variables have positive correlation with each other.



We need to quantify the correlation coefficients among the variables so that we could identify highly-correlated variables. This identification could help us avoid collinearity during variable selection process.





As shown above, many of the variables are highly correlated and it therefore would not be surprising if in the final subset of variables (best subset for prediction) for the predictive model we end up using only a few of these 30 variables.