Cancer Prediction

Group 2

10/21/2020

# Introduction

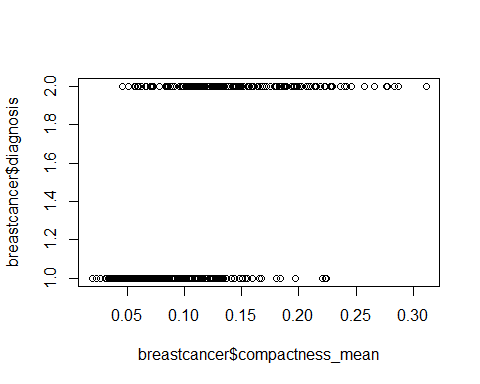
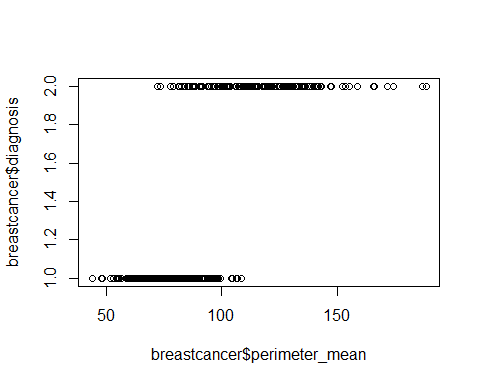
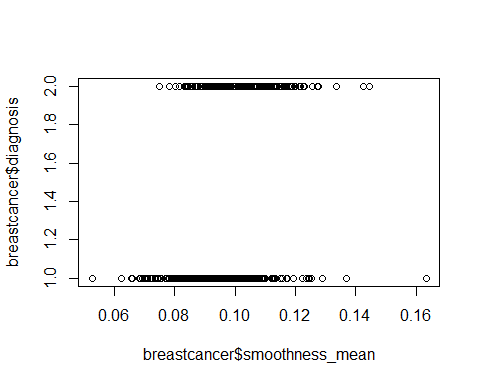
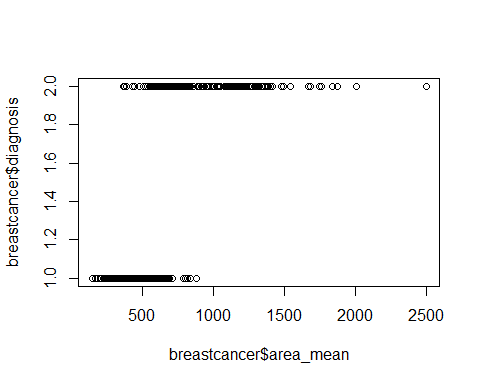
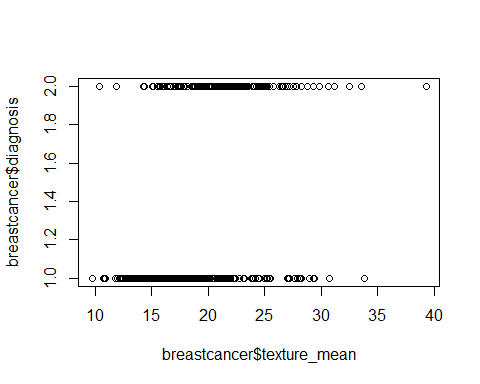
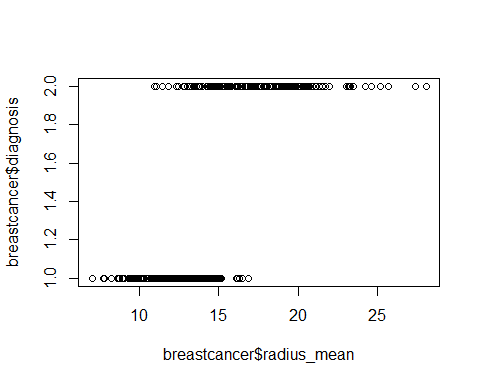
The goal of our work is to use data on tissue samples taken from cancerous breast tissue to create a logistic regression model that can be used to accurately predict if a cancer sample is malignant or benign. Our approach is to first analyze different predictor variables to check for extreme values that could influence our models accuracy. To do this we will plot predictor variables and look for outliers that could skew the variable. If need be we can then transform variables using such transformations as square roots, logarithms, and reciprocals. We will also look at the predictor variables to determine if any are highly correlated. If so, we will be able to reduce the number of predictor variables to account for overfitting our regression model. Once we have checked our predictor variables, and transformed those where we felt necessary, We then intend to build models and test their predictive capabilities.

# Exploring and Transforming the Data

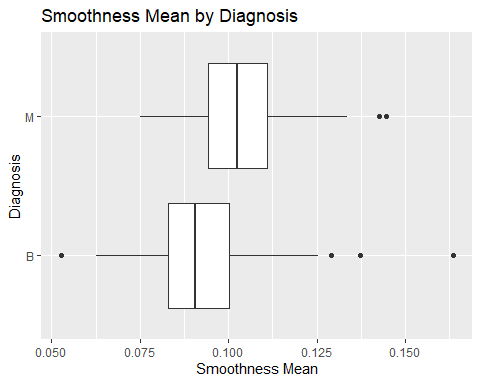
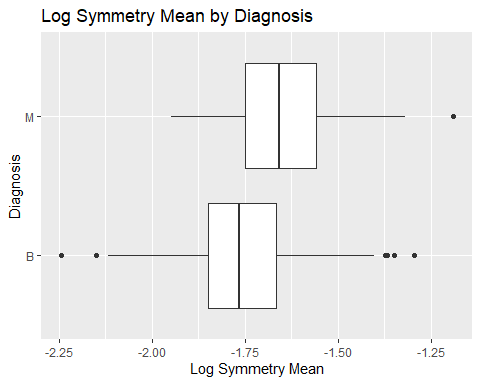
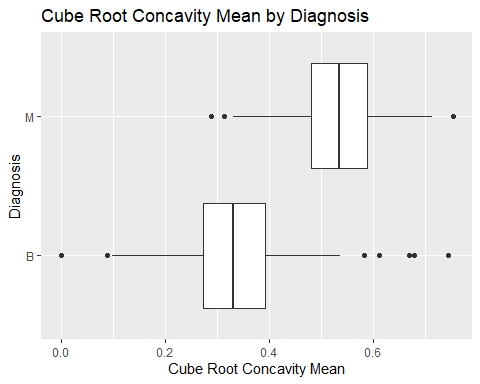
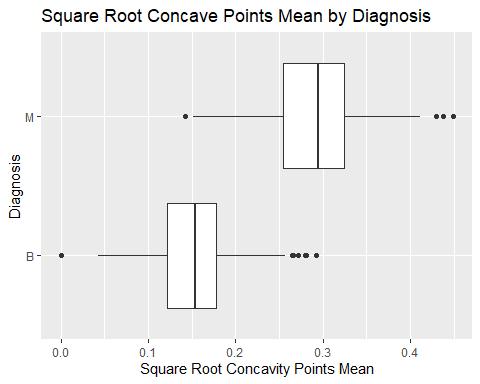
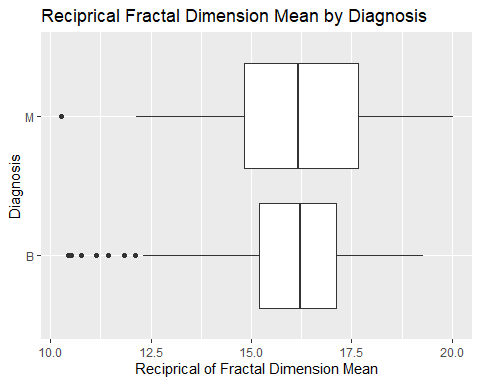
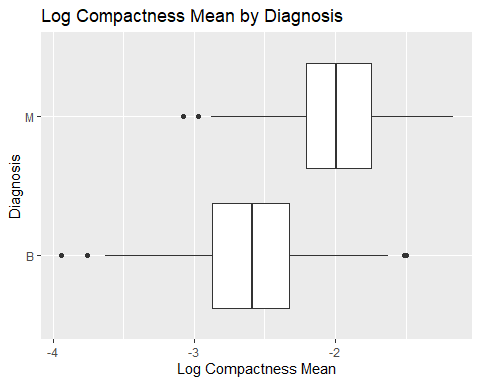
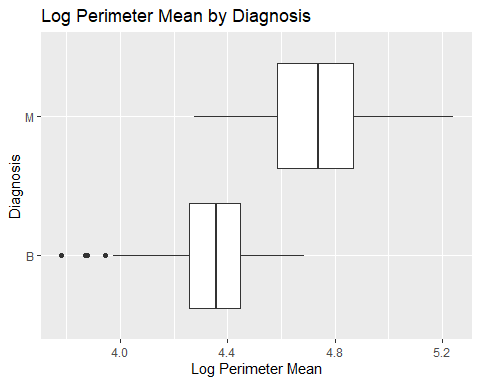
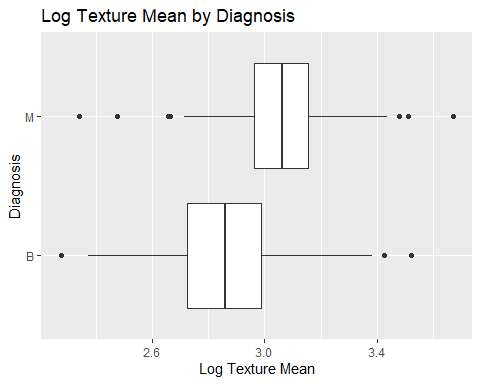
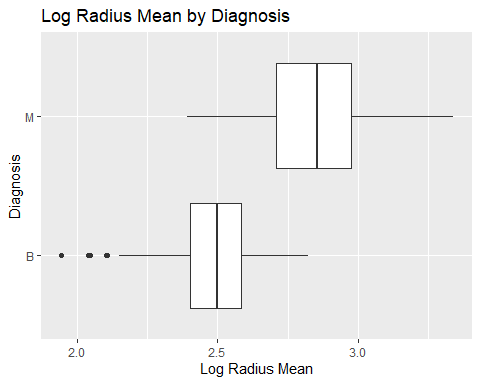
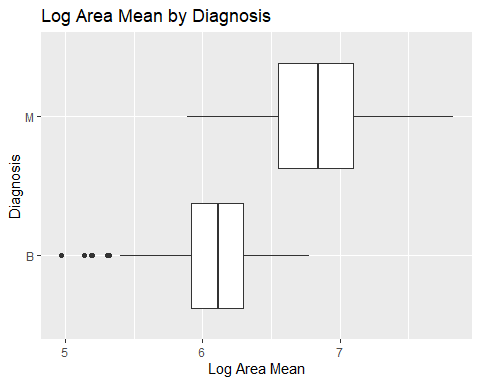
The descriptive statistics for each variable mean are displayed below:

## radius\_mean texture\_mean perimeter\_mean area\_mean   
## Min. : 6.981 Min. : 9.71 Min. : 43.79 Min. : 143.5   
## 1st Qu.:11.630 1st Qu.:16.17 1st Qu.: 74.68 1st Qu.: 415.1   
## Median :13.240 Median :18.82 Median : 85.48 Median : 541.6   
## Mean :14.044 Mean :19.31 Mean : 91.41 Mean : 648.5   
## 3rd Qu.:15.610 3rd Qu.:21.68 3rd Qu.:103.20 3rd Qu.: 761.3   
## Max. :28.110 Max. :39.28 Max. :188.50 Max. :2501.0   
## smoothness\_mean compactness\_mean concavity\_mean concave.points\_mean  
## Min. :0.05263 Min. :0.01938 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.08637 1st Qu.:0.06492 1st Qu.:0.02956 1st Qu.:0.02031   
## Median :0.09514 Median :0.09228 Median :0.06126 Median :0.03250   
## Mean :0.09617 Mean :0.10423 Mean :0.08862 Mean :0.04842   
## 3rd Qu.:0.10540 3rd Qu.:0.13040 3rd Qu.:0.13070 3rd Qu.:0.07415   
## Max. :0.16340 Max. :0.31140 Max. :0.42680 Max. :0.20120   
## symmetry\_mean fractal\_dimension\_mean  
## Min. :0.1060 Min. :0.04996   
## 1st Qu.:0.1618 1st Qu.:0.05781   
## Median :0.1788 Median :0.06166   
## Mean :0.1812 Mean :0.06296   
## 3rd Qu.:0.1966 3rd Qu.:0.06635   
## Max. :0.3040 Max. :0.09744

Scatter plots for each variable mean are displayed below:



After analyzing the summaries and the distribution of the variables we have decided to perform many transformations to account for abnormally distributed data. The transformations on each variable can be seen in the boxplots displayed below. These transformations were performed because the normal boxplots of the predictors had outliers and or highly skewed. Then the data was transformed either by a log, reciprocal, square root, and cube root transformations. These transformations allow for the distributions to be less skewed and contain less outliers making sure the data is as close to normal as possible. After performing the transformations the box plots shown below show less outliers and are less highly skewed than the original. These transformations have a more likely chance of getting better prediction models later when we conduct tests of the data. Models will also be more accurate with these changes in transformations as their are less outliers to skew the models that we will conduct later. Overall these transformations were needed to better fit models and get the data as close to normal as possible.



Based on this correlation plot we can see that the variable ‘concave.points\_mean’ and ‘concavity\_mean’ have a high positive correlation with most other predictor variable means. So, we will want to pay close attention to these variables when creating the logistic regression model to avoid multicollinearity. Additionally, we will want to examine the variables ‘radius\_mean’, ‘perimeter\_mean’, and ‘area\_mean’ because they are also highly positively correlated. 