Group 07 Assignment02

Shameka Brayton, Brad O'Hara

7/23/2017

Table of Contents



# Introduction

need to put some introductory words here describing what we will be doing

# Data Gathering

For this assignment my team and I decided to use the wisconsin breast cancer research data found on the UCI website.  
- To begin the process, we pulled the data directly from the URL into a data frame.  
- Then we replaced the column headers wich were not included in the data, but we were able to pull from the data dictionary made available by the UCI site.

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:reshape2':  
##   
## dcast, melt

Attributes for the dataset were given as follows:

## Data Featrues

Sample code number id number Clump Thickness  
Uniformity of Cell Size  
Uniformity of Cell Shape  
Marginal Adhesion  
Single Epithelial Cell Size  
Bare Nuclei  
Bland Chromatin  
Normal Nucleoli  
Mitoses  
Classfor benignfor malignant

# Data Cleansing

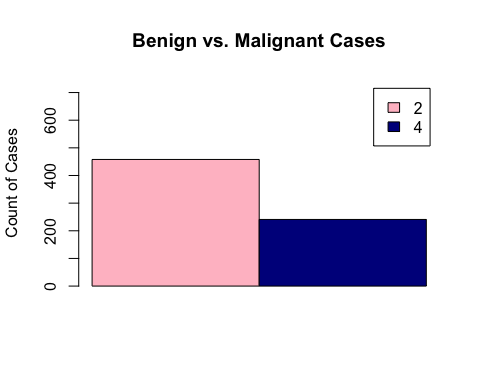
here we had to normalize 16 rows of data that contained NA's by using the DMwR2 funciton "centralImputation" we were able to use the median to fill in for the NA entries within our dataset.

Top 3 Rows with Missing Data

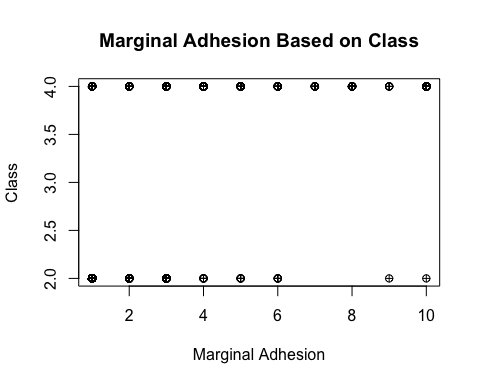
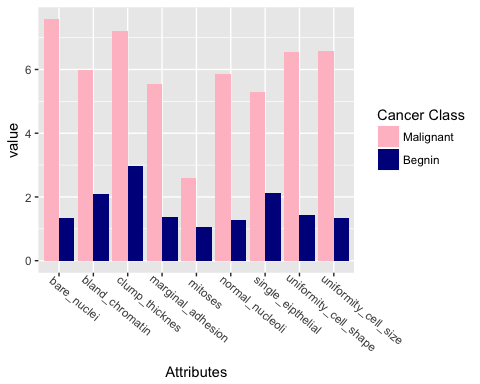
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patientID | clump\_thicknes | uniformity\_cell\_size | uniformity\_cell\_shape | marginal\_adhesion | single\_eipthelial | bare\_nuclei | bland\_chromatin | normal\_nucleoli | mitoses | class |
| 1057013 | 8 | 4 | 5 | 1 | 2 | NA | 7 | 3 | 1 | 4 |
| 1096800 | 6 | 6 | 6 | 9 | 6 | NA | 7 | 8 | 1 | 2 |
| 1183246 | 1 | 1 | 1 | 1 | 1 | NA | 2 | 1 | 1 | 2 |

Same 3 rows, with NA's cleansed

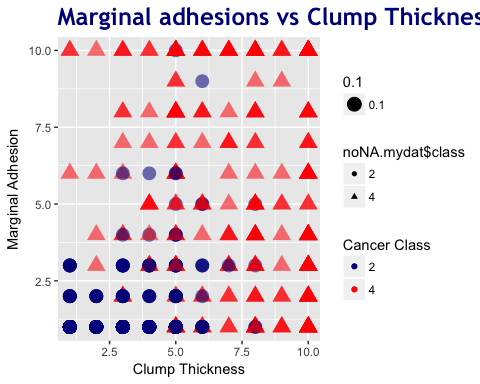
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patientID | clump\_thicknes | uniformity\_cell\_size | uniformity\_cell\_shape | marginal\_adhesion | single\_eipthelial | bare\_nuclei | bland\_chromatin | normal\_nucleoli | mitoses | class |
| 1057013 | 8 | 4 | 5 | 1 | 2 | 1 | 7 | 3 | 1 | 4 |
| 1096800 | 6 | 6 | 6 | 9 | 6 | 1 | 7 | 8 | 1 | 2 |
| 1183246 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |

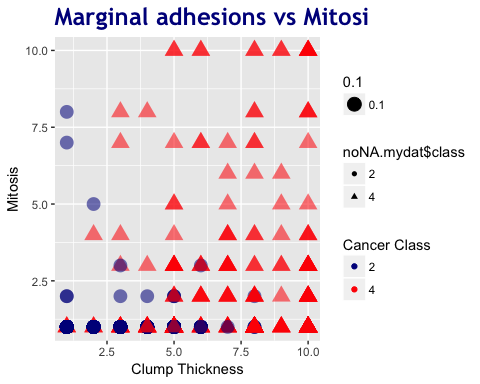
here we show the relationship between Benign and Malignant cases contained in our data. 

##   
## Call:  
## glm(formula = as.factor(noNA.mydat$class) ~ marginal\_adhesion +   
## clump\_thicknes + uniformity\_cell\_size + uniformity\_cell\_shape +   
## single\_eipthelial + bare\_nuclei + bland\_chromatin + normal\_nucleoli +   
## mitoses, family = binomial, data = noNA.mydat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3310 -0.1269 -0.0659 0.0273 2.3954   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.71454 1.07434 -9.042 < 2e-16 \*\*\*  
## marginal\_adhesion 0.23762 0.11672 2.036 0.04176 \*   
## clump\_thicknes 0.53465 0.13488 3.964 7.38e-05 \*\*\*  
## uniformity\_cell\_size 0.01128 0.19321 0.058 0.95344   
## uniformity\_cell\_shape 0.32377 0.21390 1.514 0.13011   
## single\_eipthelial 0.05832 0.15255 0.382 0.70222   
## bare\_nuclei 0.42816 0.09051 4.730 2.24e-06 \*\*\*  
## bland\_chromatin 0.41213 0.15693 2.626 0.00864 \*\*   
## normal\_nucleoli 0.15824 0.10414 1.519 0.12864   
## mitoses 0.53584 0.30698 1.746 0.08089 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 900.53 on 698 degrees of freedom  
## Residual deviance: 113.09 on 689 degrees of freedom  
## AIC: 133.09  
##   
## Number of Fisher Scoring iterations: 8

##### 



 ```