Breast Cancer Wisconsin

Rose Hazenberg

2021-09-28

Contents

1 Introduction

The research question is: Is it possible to reliably predict the cancer stage based on the uniformity of cell size/shape using machine learning?

1.1 Data description

The data obtain information about breast cancer in Wisconsin. This was provided by Dr. William H. Wolberg of the University of Wisconsin Hospitals.

1.2 Data set information

The samples are clinical cases of Dr. Wolberg which are periodically classified and are ordered in a chronological grouping. The grouping is as followed:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated database on 15 July 1992)

1.3 Attribute information

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

2 Exploratory Data Analysis (EDA) of breast cancer

Before we visualize the data we need to load the data. After the following code, the data is available to use.

```r
## Define the data file
datafile <- "data/breast-cancer-wisconsin.data"

## Load the dataset
breastcancer <- read.table(datafile, sep = ",", header = FALSE,
                           na.strings = "?")

## Changed the column names
colnames(breastcancer) <- c("ID", "clump_thickness", "uniformity_of_cell_size",
                            "uniformity_of_cell_shape", "marginal_adhesion",
                            "single_epithelial_cell_size", "bare_nuclei",
                            "bland_chromatin", "normal_nucleoli", "mitoses",
                            "class")

## Create new variables for the column class
breastcancer <- breastcancer %>%
    mutate(class = factor(class, labels = c("Benign", "Malignant"),
                          levels = c(2, 4)))

## Inspect the data
str(breastcancer)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ ID                         : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561
##  $ clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ uniformity_of_cell_size    : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ uniformity_of_cell_shape   : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
##  $ bare_nuclei                : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses                    : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class                      : Factor w/ 2 levels "Benign","Malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

```r
## Show the first 6 rows of the data
head(breastcancer)
```

```
##        ID clump_thickness uniformity_of_cell_size uniformity_of_cell_shape
## 1 1000025               5                       1                        1
## 2 1002945               5                       4                        4
## 3 1015425               3                       1                        1
## 4 1016277               6                       8                        8
## 5 1017023               4                       1                        1
## 6 1017122               8                      10                       10
##   marginal_adhesion single_epithelial_cell_size bare_nuclei bland_chromatin
## 1                 1                           2           1               3
## 2                 5                           7          10               3
## 3                 1                           2           2               3
## 4                 1                           3           4               3
## 5                 3                           2           1               3
## 6                 8                           7          10               9
##   normal_nucleoli mitoses     class
## 1               1       1    Benign
## 2               2       1    Benign
## 3               1       1    Benign
## 4               7       1    Benign
## 5               1       1    Benign
## 6               7       1 Malignant
```

```r
## Show the summary per column
summary(breastcancer)
```

```
##        ID           clump_thickness  uniformity_of_cell_size
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000         
##  1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000         
##  Median : 1171710   Median : 4.000   Median : 1.000         
##  Mean   : 1071704   Mean   : 4.418   Mean   : 3.134         
##  3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000         
##  Max.   :13454352   Max.   :10.000   Max.   :10.000         
##                                                             
##  uniformity_of_cell_shape marginal_adhesion single_epithelial_cell_size
##  Min.   : 1.000           Min.   : 1.000    Min.   : 1.000             
##  1st Qu.: 1.000           1st Qu.: 1.000    1st Qu.: 2.000             
##  Median : 1.000           Median : 1.000    Median : 2.000             
##  Mean   : 3.207           Mean   : 2.807    Mean   : 3.216             
##  3rd Qu.: 5.000           3rd Qu.: 4.000    3rd Qu.: 4.000             
##  Max.   :10.000           Max.   :10.000    Max.   :10.000             
##                                                                        
##   bare_nuclei     bland_chromatin  normal_nucleoli     mitoses     
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000  
##  1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000  
##  Median : 1.000   Median : 3.000   Median : 1.000   Median : 1.000  
```

Table 1: The number of NA's per column

	Number of NA's
ID	0
clump_thickness	0
uniformity_of_cell_size	0
uniformity_of_cell_shape	0
marginal_adhesion	0
single_epithelial_cell_size	0
bare_nuclei	16
bland_chromatin	0
normal_nucleoli	0
mitoses	0
class	0

```
##  Mean   : 3.545   Mean   : 3.438   Mean   : 2.867   Mean   : 1.589
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##  NA's   :16
##        class
##  Benign   :458
##  Malignant:241
##
##
##
##
##
```

The outcome of the summary shows that every column has the same minimum and maximum except the id and the class but these columns aren't properties. The 1st Qu, median, mean and 3rd Qu are different for every column but they all lie close to each other. At the minimum and maximum, you can see that the data only consist of the numbers 1 to 10.

2.1 Missing data

To check if there is missing data, a table is generated with the number of NA's per column with the following code.

```
## Create table of NA's
kable(colSums(is.na(breastcancer)), col.names = "Number of NA's",
      caption = "The number of NA's per column")
```

Table 1 shows the number of NA's in the dataset of breast cancer. It shows that there are 16 NA's in de column bare nuclei, the other columns have 0 NA's. For the research question, there will be looked at the uniformity of cell size/shape and not at the column bare nuclei. Since there are only 16 NA's and in the column that isn't relevant so it doesn't affect the research question. To deal with these missing values column bare nuclei can be dropped. For the visualizations now it doesn't impact the data so these values aren't removed.

Table 2: Head of the ’lengthens’ data

ID	class	property	value
1000025	Benign	clump_thickness	5
1000025	Benign	uniformity_of_cell_size	1
1000025	Benign	uniformity_of_cell_shape	1
1000025	Benign	marginal_adhesion	1
1000025	Benign	single_epithelial_cell_size	2
1000025	Benign	bare_nuclei	1

2.2 Variation/distribution

Here is the variation within the data and the distribution of the data examined.

```r
## Tidying the data by using pivot_longer
breast_long <- pivot_longer(data = breastcancer,
                            cols = -c("ID", "class"),
                            names_to = "property",
                            values_to = "value")

## Show the first 6 rows of the breastcancer dataset after pivot_longer
kable(head(breast_long), digits = 6, caption = "Head of the 'lengthens' data")
```

In table 2 the number of rows is increased and the number of columns is decreased compared to the original loaded data. With the new form of the data it can be visualized easily per property.

2.2.1 Histogram

One of the visualization to examine the variation and the distribution is a histogram.

```r
## Plot the histogram
ggplot(data = breast_long, mapping = aes(x = property, y = value)) +
  geom_histogram(stat = 'identity', aes(color = class, fill = class)) +
  scale_fill_manual(values = c("steelblue", "plum1"),
                    aesthetics = c("color", "fill")) +
  labs(title="Histogram",
       subtitle="Total count per property per stage",
       x="Property",
       y="Total") +
  coord_flip() +
  theme_classic()
```

Figure 1 shows the total count of the properties per stage in a histogram. The plot shows the properties for the class benign and malignant. The variance between the total of the properties per class lie close to each other except for the properties mitoses, here is the total the lowest, and clump thickness, here is the total the highest. Figure 1 shows that the distribution of the classes differ from each other because the class benign has a smaller bar than the class malignant and so a lower count for benign. The data consist of the number 1 to 10 and this says something about how benign or malignant a property is for the id of the person. So with this figure 1 you can’t say which class belongs to the person only what the total count of a property is for a class. The class benign has mainly values of 1 so this explains why the bar of the class benign per property is smaller. And the class malignant has mainly values of 10 so this bar is longer but the class can also have values of 1. These values can’t say which attribute is more informative in the model process because of the difference in values of the properties. But it can say that the class malignant has higher values than the class benign because the total count and the bar is bigger.

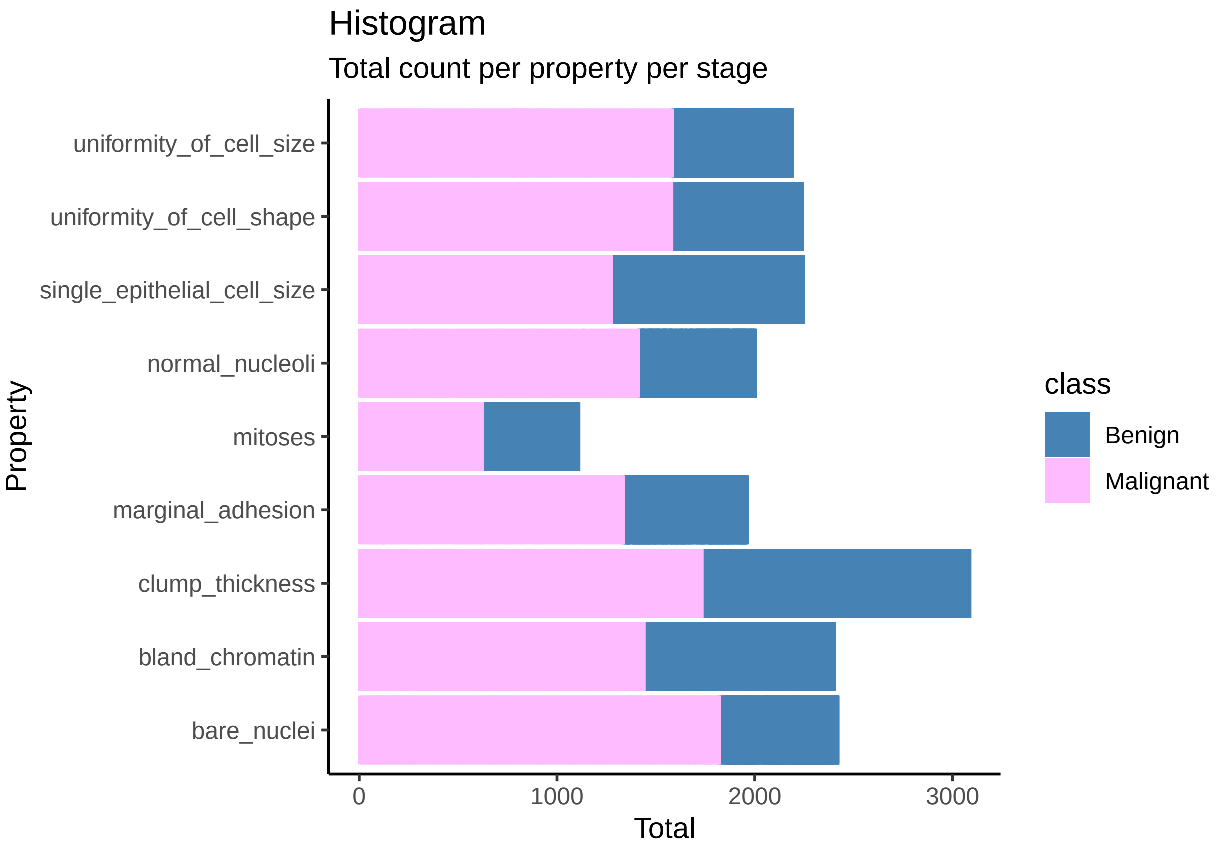

Figure 1: Histogram with the count of the properties per stage of breast cancer

2.2.2 Density plot

Another visualization to examine the data is a density plot. Which is made and describe below.

```r
## Plot the density plot using ggplot
ggplot(data = breast_long, mapping = aes(value, color = property)) +
  geom_density() +
  labs(title="Density plot",
       subtitle="The density of the value",
       x = "Value",
       y = "Density") +
  theme_minimal() +
  scale_x_continuous(breaks = c(1:10))
```

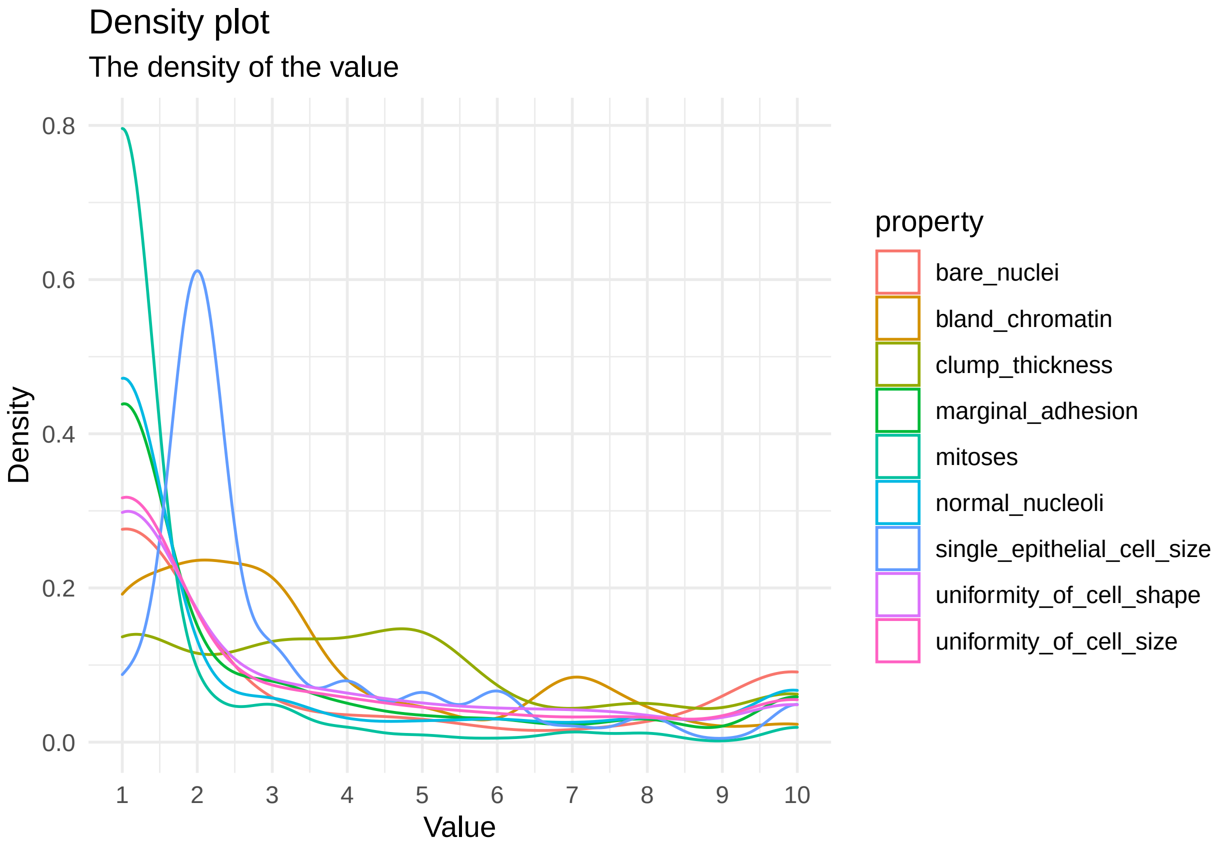

Figure 2: Density plot with the probability density of the value per property

The density plot is shown in figure 2 which shows the density of the values per property. The density plot shows the shape of the distribution of the values per property. The property mitoses shows by the value 1 a density of 0.8 which is the highest of them all. Single epithelial cell size is the only property with a curve of this distribution. The other properties have a density that runs in waves. The variation in density between the property lies mainly between the value 1 to 4 because here is the density higher and the density decreases from value 1 until around value 4. After value 4 the lines of the properties run almost around the same density. Thus at higher values, the values approximately occur equal per property.

2.2.3 Boxplot

The last visualization is a boxplot. A boxplot shows the five-number summary of the data: including the minimum score, lower (Q1) quartile, median, upper (Q3) quartile, and maximum score.

```r
## Plot the boxplot with ggplot
ggplot(data = breast_long, mapping = aes(property, value, color = class,
                                         fill = class)) +
  geom_boxplot(aes(outlier.colour = class), alpha = 0.6) +
  scale_colour_manual(name = "class", values = c("steelblue", "plum1"),
                      aesthetics = c("color", "fill")) +
  labs(title="Boxplot",
       subtitle="Property per stage of breastcancer",
       x="Property",
       y="Value") +
  coord_flip() +
  theme_classic() +
  scale_y_continuous(breaks = c(1:10))
```

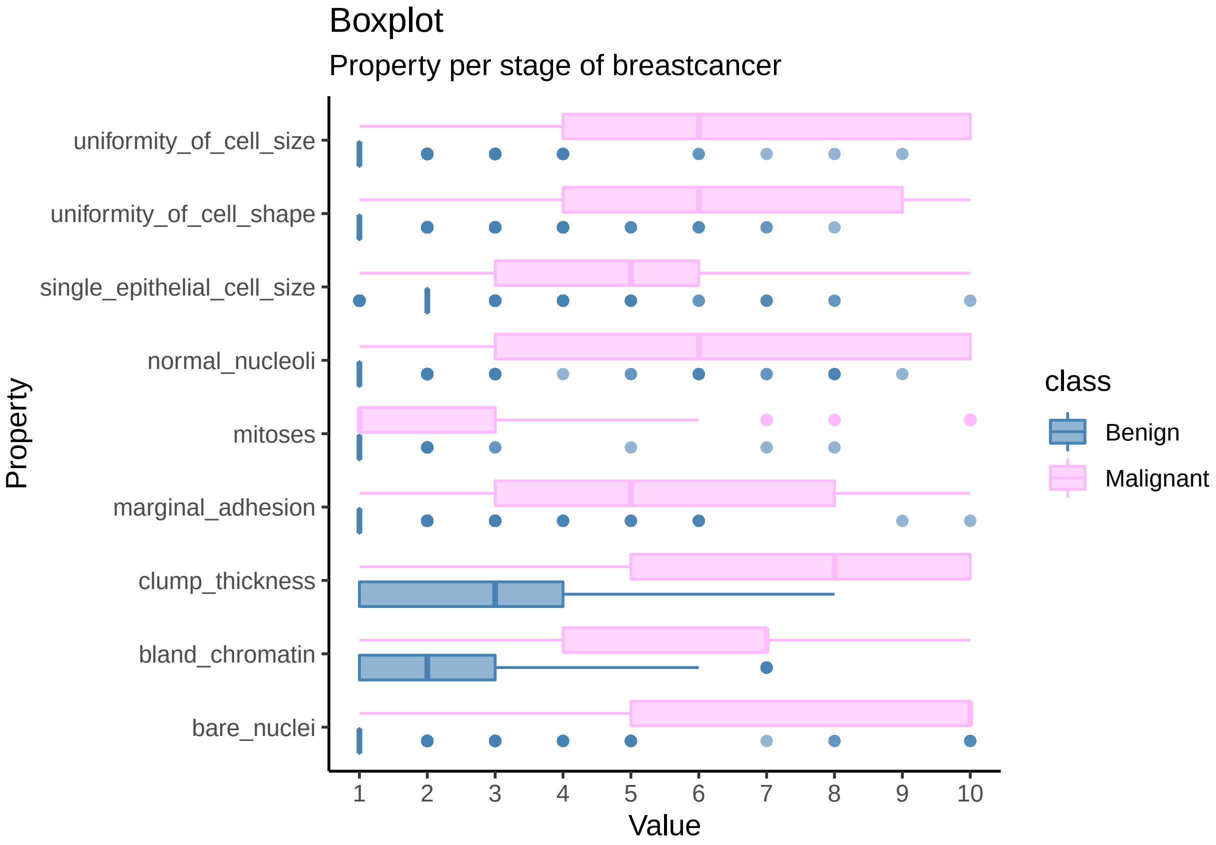

Figure 3: Boxplot comparing for all properties per class

Figure 3 shows multiple boxplots with the distribution of the values per property divided into the two classes with outliers. For every property of the class malignant is there a boxplot but for the class benign there are only two that can be seen the other boxplots are just a line mainly at value 1. The boxplots of class benign consist of outliers except for the properties clump thickness and bland chromatin. This is because benign consists mainly of the value 1 and the other values are thus outliers. The boxplots of the class malignant lie mainly to the right and are longer and they don't have any outliers except the property mitosis. But just like in figure 2 shows that mitoses consist mainly of the value 1 until 3 so that is why this boxplot is more to the left and has outliers. For the class malignant the outliers can be removed there are outliers by mitoses. But for the class benign could the outliers be informative for the class but also for the properties because there are only two bigger boxplots and the other properties consist of a small boxplot and mainly outliers.

2.3 Class distribution

2.3.1 Bar plot

To check the class distribution, so if the classes are evenly or unevenly represented a bar plot is used.

```r
## Plot the bar plot using ggplot
ggplot(data = breast_long, mapping = aes(class, value)) +
  geom_bar(stat = "identity", aes(color = class, fill = class)) +
  scale_fill_manual(values=c("steelblue", "plum1"),
                    aesthetics = c("color", "fill")) +
  labs(title="Bar plot",
       subtitle="Total count per stage of breastcancer",
       x="Class",
       y="Total") +
  theme_minimal()
```

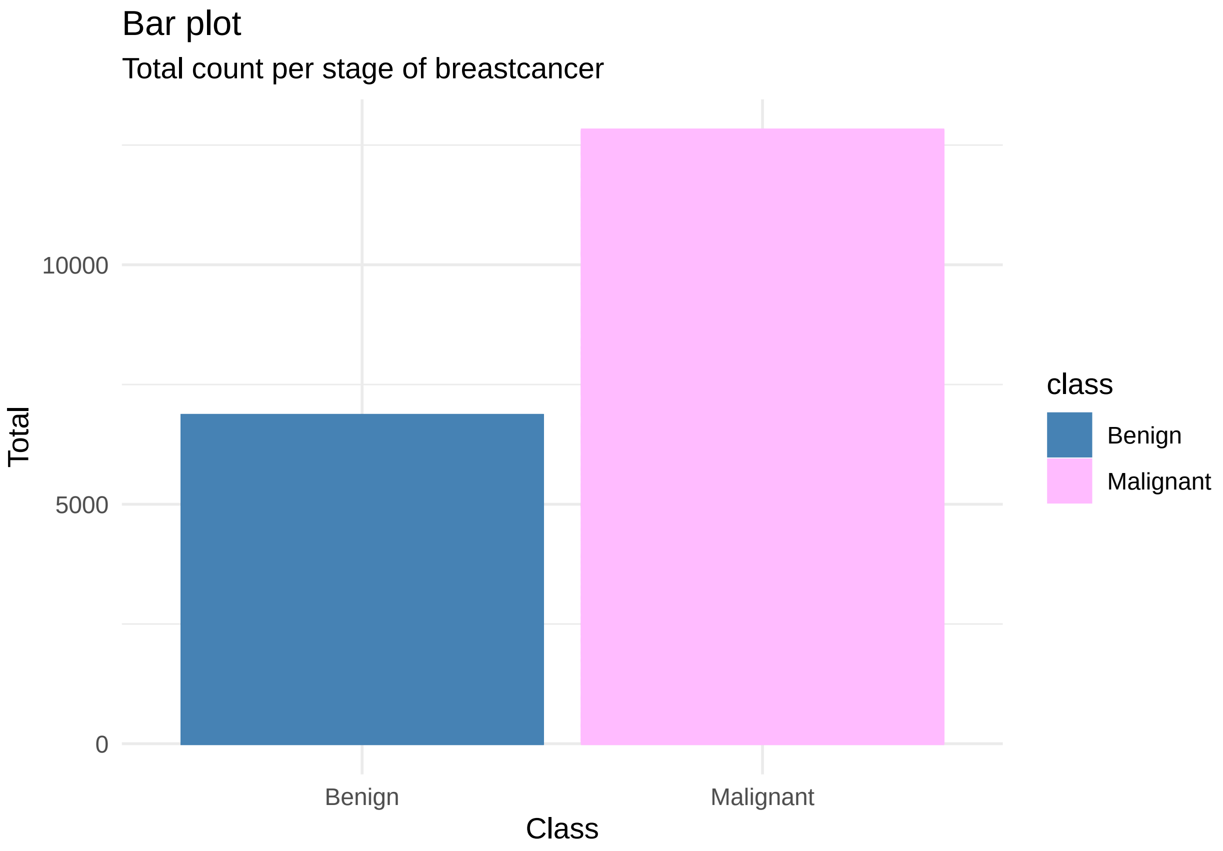

Figure 4: Bar plot of the total number of values per class

The bar plot in figure 4 shows the total number of values per class. Here you can see that the classes are unevenly represented, the total is higher for the class malignant. But this can be, as described in figure 1, that the class benign mainly consist of low values such as 1 and that the class malignant mainly consists of high values such as 10. The figure shows the total of these values. So this could explain why the bar for the class malignant is higher. This doesn't say that there is more malignant than benign.

Table 3: Total number of the class

Class	Total
Benign	458
Malignant	241

2.3.2 Table

```r
## Create table
kable(table(breastcancer$class), col.names = c("Class", "Total"),
      caption = "Total number of the class")
```

Table 3 shows the total number of benign and malignant. Figure 4 shows that the total of malignant is higher than benign but in table 3 you can see that the total of the class benign is more. This is because table 3 shows the total number of rows with the class and figure 4 shows the total count of the values 1 to 10 for each class. So in table 3 you can see that there isn't more malignant than benign. So here are the classes also unevenly represented but then the other way around.

2.4 Correlation

2.4.1 Scatter plot

```
## Plot the scatter plot
ggplot(data = breast_long, mapping = aes(x = property, y = value)) +
  geom_point(aes(color = class), alpha = 0.7) +
  geom_smooth(aes(group = class), method = "lm") +
  scale_fill_manual(values = c("steelblue", "plum1"),
                    aesthetics = c("color", "fill")) +
  labs(title="Scatter plot",
       subtitle="Total count per stage of breastcancer",
       x="Property",
       y="Value") +
  theme_minimal() +
  theme(axis.text = element_text(angle = 90)) +
  facet_wrap(. ~ class) +
  scale_y_continuous(breaks = c(1:10))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

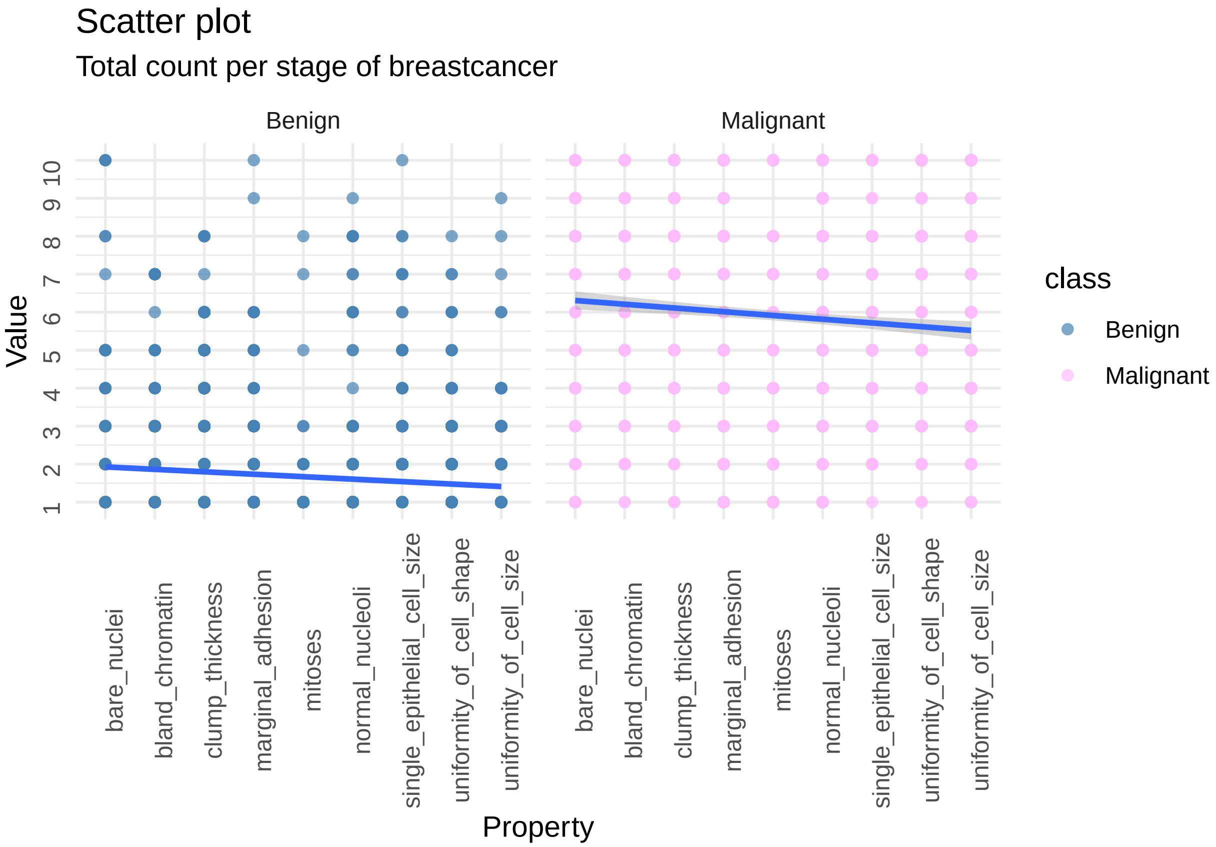

Figure 5: Scatter plot with a predicted linear regression.

The scatter plot in figure 5 shows the total count of the values, the dots, per property for each class. The plot contains a linear regression of the blue lines. In the scatter plot of benign can be seen that it doesn't have all values per property and that the most values lie between 1 and 3 because these dots are darker. The line runs from 2 until almost 1 and it decreased so it has a negative relationship and thus the correlation. For the class malignant there are dots everywhere except at value 9 by the property mitoses. Just like in the plot of benign the line decreased but only higher around 6.5 until 5.5 and it has also a negative relationship.

2.4.2 Heatmap

```
## Plot the heatmap using geom_title
ggplot(data = breast_long, mapping = aes(x = property, y = class,
                                         fill = value)) +
  geom_tile() +
  labs(title="Heatmap",
       x="",
       y="") +
  theme_minimal() +
  theme(axis.text = element_text(angle = 90))
```

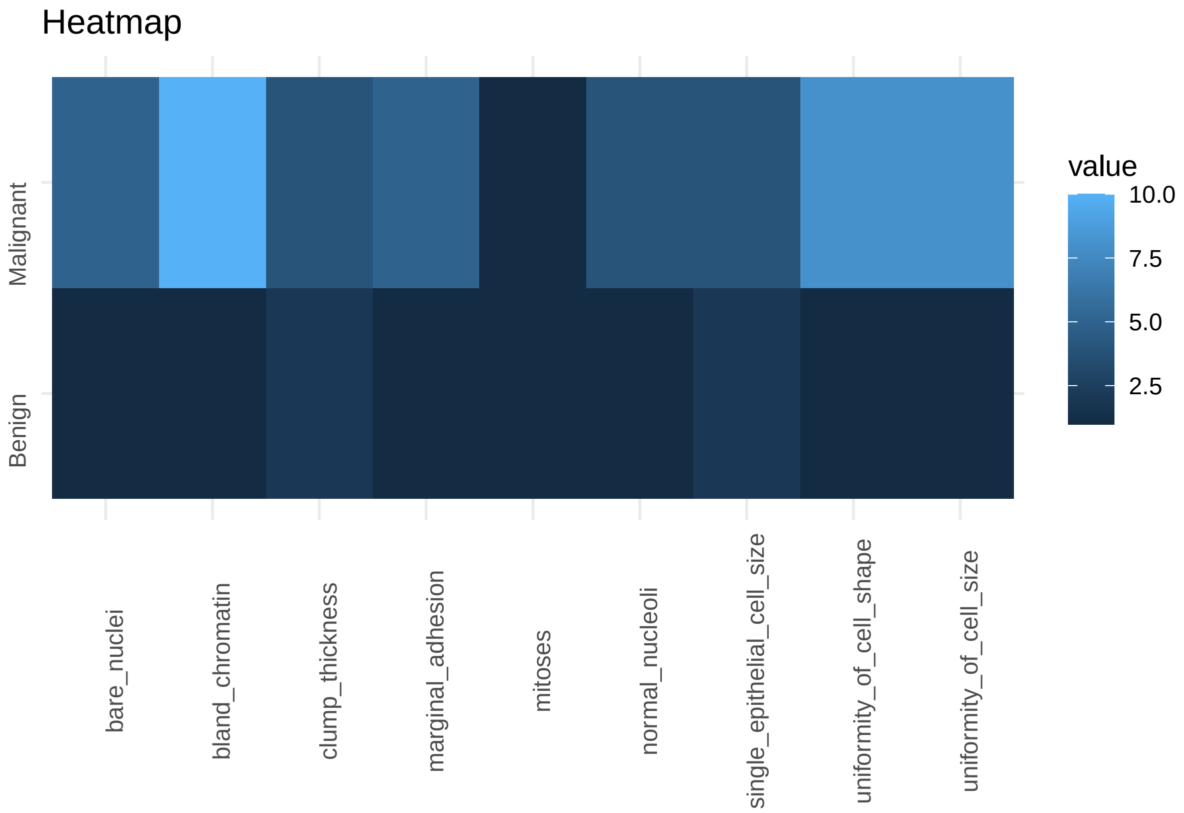

Figure 6: Heatmap pairwise corrrelation of the values

Figure 6 shows the correlation of the values for the properties and the classes. The class benign has mainly low values which show that they are low correlated. Except the properties clump thickness and single epithelial cell size where the values are higher so they are higher correlation than the other properties. Almost every property is correlated differently for the class malignant. The property bland chromatin is highly correlated and mitosis is low correlated compared to the other properties which lie in between.

2.5 Clustering

2.5.1 kMeans clustering

The next two figures are clusters based on two properties the uniformity of cell size and shape.

```r
kmeans.cluster <- kmeans(is.na(breastcancer), centers = 2)

clusters <- as.factor(kmeans.cluster$cluster)

ggplot(breastcancer, aes(uniformity_of_cell_size, class, color = clusters)) +
  geom_point(size = 0.5, position = position_jitter(height = 0.3, width = 0.3)) +
  labs(title="Clusters of Uniformity of cell size per class",
       x="Uniformity of cell size",
       y="Class") +
  theme_minimal() +
  scale_x_continuous(breaks = c(1:10))
```

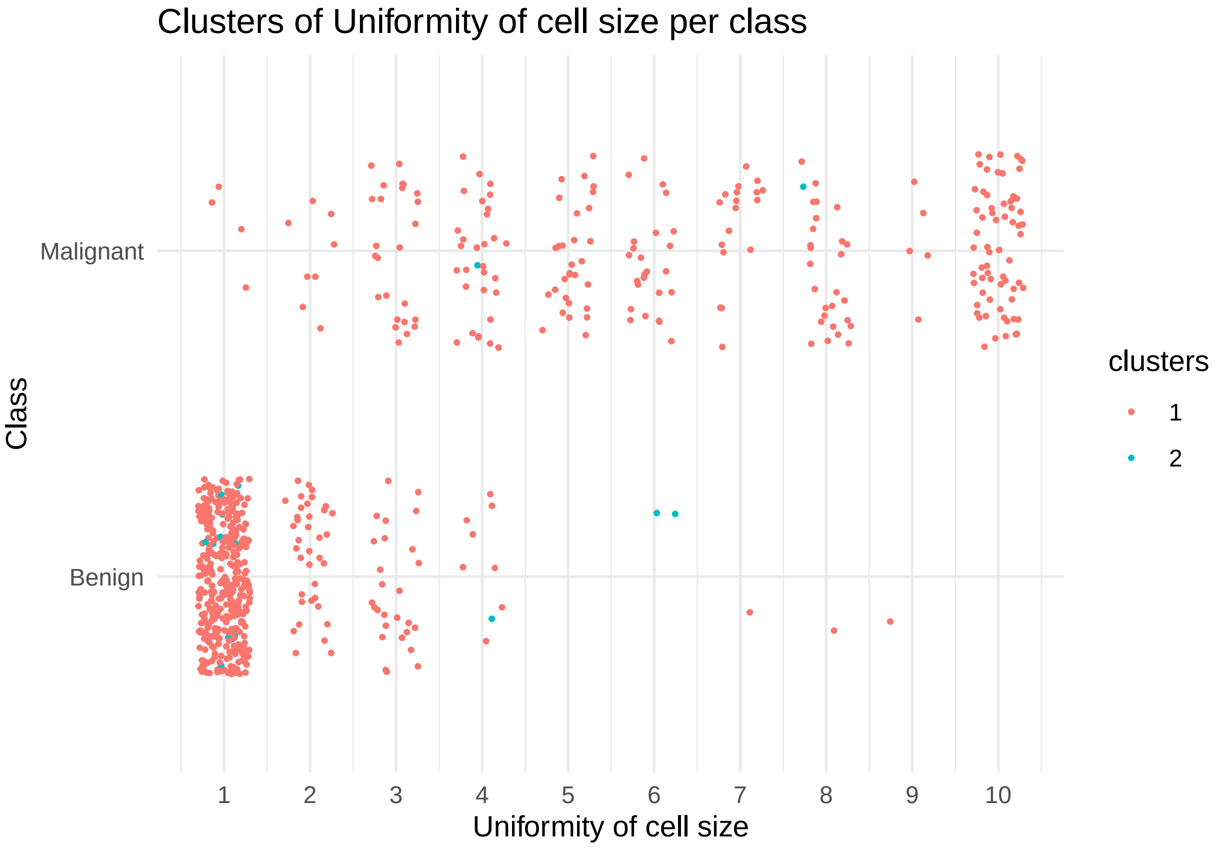

Figure 7: kMeans cluster of uniformity of cell size per class

Figure ?? shows a kMeans cluster of the property uniformity of cell size for both classes. It shows that there is a cluster for the class benign at value 1, and there is small cluster at values 2 and 3. The other blue dots are scattered between values 4 and 9. For the class malignant is there a cluster at value 10 and smaller clusters at value 3 until 8. For this class are the blue dots are more scattered over the values. The cluster for both classes can be explained that benign has more 1 values and malignant has more 10 values. By both classes are the pink dots scattered in the plot these are small clusters.

```
ggplot(breastcancer, aes(uniformity_of_cell_shape, class, color = clusters)) +
  geom_point(size = 0.5, position = position_jitter(height = 0.3, width = 0.3)) +
  labs(title="Clusters of Uniformity of cell shape per class",
       x="Uniformity of cell shape",
       y="Class") +
  theme_minimal() +
  scale_x_continuous(breaks = c(1:10))
```

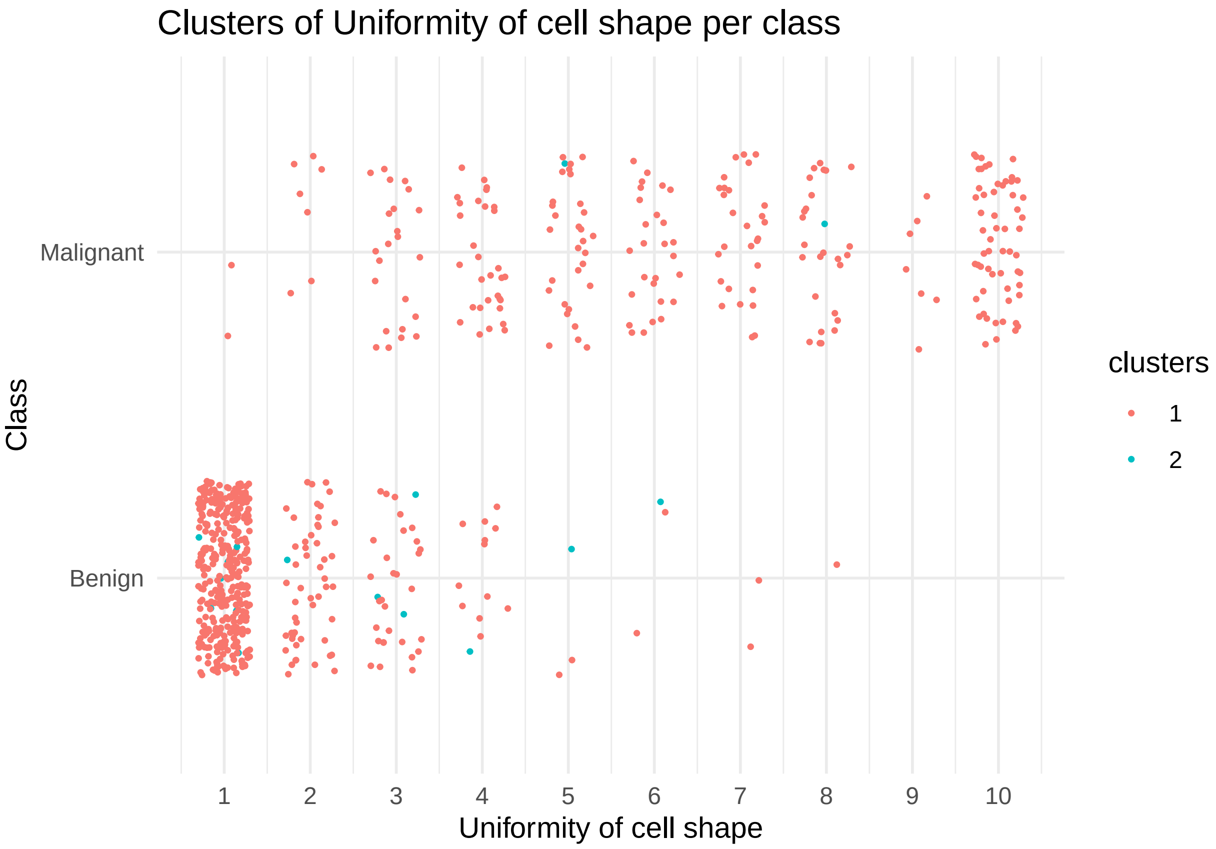

Figure 8: kMeans cluster of uniformity of cell shape per class

This figure 8 looks almost the same as the plot in figure **??** but here are the blue dots more scattered between the values 3 until 8 and there are more pink clusters. The class malignant has in one eye the same shape of clusters as in figure 8 and it also has two pink clusters.

2.5.2 Principal Components Analysis (PCA)

```
## Impute dataset with PCA
bc.comp <- imputePCA(breastcancer[,2:10], graph = FALSE)
bc.pca <- PCA(bc.comp$completeObs, scale.unit = TRUE, ncp = 5, graph = FALSE)

## Plot the PCA
fviz_pca_var(bc.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, alpha.var = 0.7)
```

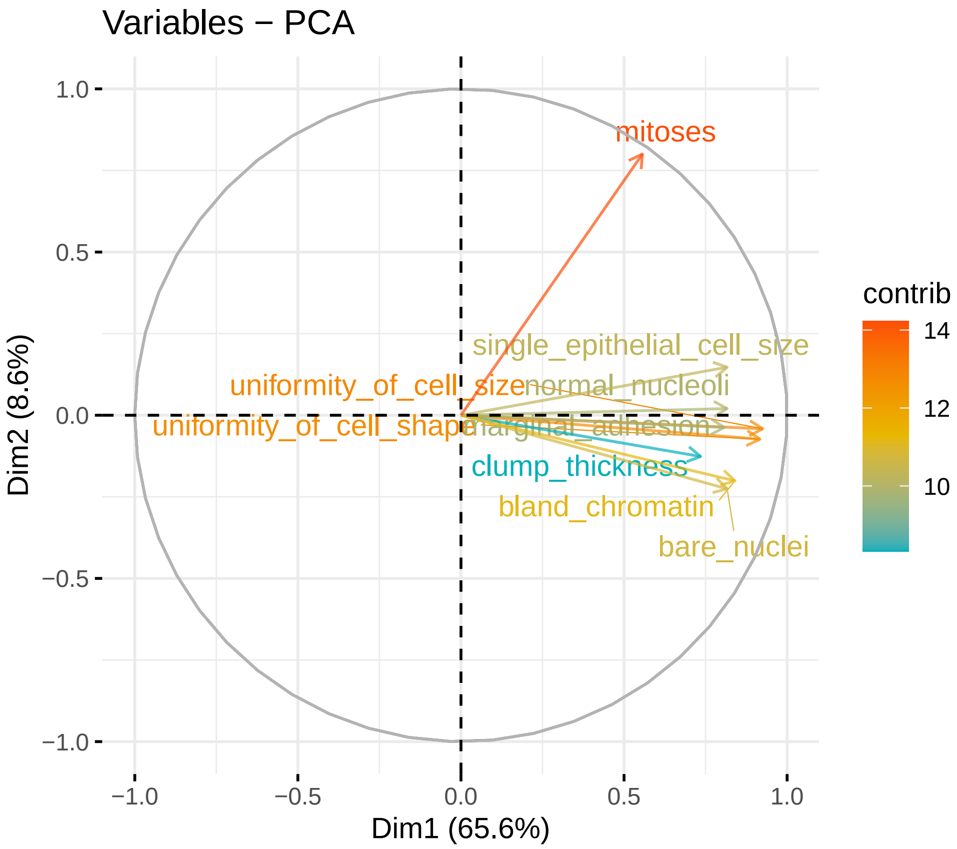

Figure 9: Correlation plot with the contribution of variables

Figure 9 shows a plot of the contribution per dimension which has two dimensions. The contribution show a correlation between variables. In the plot, there aren't properties that have an arrow to the left and aren't negative. But some properties point to the bottom right so they are negative for dim2 and positive for dim1. The properties/variables mitoses, uniformity_of_cell_size and uniformity_of_cell_shape contribute the most to the dimensions 1 and 2 because these properties have a high contrib value of 13/14. The property clump thickness shows the least contribute variable because the variable is colored blue. The other variables lie in between.