EDA.rmd

Carlo van Buiten

9/15/2021

Research question

Is cell uniformity (shape and size) a good predictor for the malignancy of breast cancer?

In this dataset all attributes are grades ranging from 1 to 10, with 10 being the most severe. Each row is a unique case of breast-cancer. For specific percentages etc. refer to the user manual below:

Attributes definitions: https://www.rai-light.com/docs/BCD_User_Manual_v01.pdf

I hypothesize that it makes sense for a lower cell uniformity to coincide with a higher severity of (breast) cancer, since one the defining aspects of cancer is the erratic way the cells grow and multiply.

Let us start by taking a look at the data distribution.

```
data <- read.csv("breast-cancer-wisconsin.data", na.strings="?")
colnames(data) <- c("id", "Clump_Thickness", "Cell_Size_Uniformity", "Cell_Shape_Uniformity", "Marginal
                    "Single_Epithelial_Cell_Size", "Bare_Nuclei", "Bland_Chromatin", "Normal_Nucleoli",
                    "Class")

# head(data, 10)
summary(data)
```

```
##        id           Clump_Thickness  Cell_Size_Uniformity Cell_Shape_Uniformity
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000       Min.   : 1.000
##  1st Qu.:  870258   1st Qu.: 2.000   1st Qu.: 1.000       1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000       Median : 1.000
##  Mean   : 1071807   Mean   : 4.417   Mean   : 3.138       Mean   : 3.211
##  3rd Qu.: 1238354   3rd Qu.: 6.000   3rd Qu.: 5.000       3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000       Max.   :10.000
##
##  Marginal_Adhesion Single_Epithelial_Cell_Size  Bare_Nuclei
##  Min.   : 1.000    Min.   : 1.000              Min.   : 1.000
##  1st Qu.: 1.000    1st Qu.: 2.000              1st Qu.: 1.000
##  Median : 1.000    Median : 2.000              Median : 1.000
##  Mean   : 2.809    Mean   : 3.218              Mean   : 3.548
##  3rd Qu.: 4.000    3rd Qu.: 4.000              3rd Qu.: 6.000
##  Max.   :10.000    Max.   :10.000              Max.   :10.000
##                                                NA's   :16
##  Bland_Chromatin  Normal_Nucleoli    Mitoses          Class
##  Min.   : 1.000   Min.   : 1.00   Min.   : 1.00   Min.   :2.000
##  1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.: 1.00   1st Qu.:2.000
##  Median : 3.000   Median : 1.00   Median : 1.00   Median :2.000
##  Mean   : 3.438   Mean   : 2.87   Mean   : 1.59   Mean   :2.691
##  3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 1.00   3rd Qu.:4.000
##  Max.   :10.000   Max.   :10.00   Max.   :10.00   Max.   :4.000
##
```

We can see the data does not follow a normal distribution, this becomes apparent when you compare the column means to the medians.

Let us take a look at the highest correlating features by only taking the data columns that correlate with malignancy for more than 80%.

```
cor_data <- cor(data, use = "complete.obs")
cor_data <- as.data.frame(cor_data)
cor_cols <- names(cor_data)[which(cor_data$Class > 0.8)]
cor_cols
```

```
## [1] "Cell_Size_Uniformity"  "Cell_Shape_Uniformity" "Bare_Nuclei"
```

```
## [4] "Class"
```

There are 3 attributes that correlate with malignancy for more than 80%, 2 of which are pertaining to cell uniformity. This bodes well for the hypothesis that cell uniformity could potentially be a good predictor of malignant breast cancer.

Let us look at what the spreads of these attributes look like with regards to the 2 classes, where B is benign and M is malignant.

```
plot_data <- data
plot_data$Class[plot_data$Class == 2] <- "B"
plot_data$Class[plot_data$Class == 4] <- "M"

ggplot(data=plot_data, mapping = aes(x = Class, y = Cell_Size_Uniformity)) +
  geom_jitter(width = 0.05, height = 0.3, col="red", alpha = 0.3) +
  labs(y="Cell Size Uniformity",
       caption="Figure 1: Cell size uniformity spread per class.") +
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

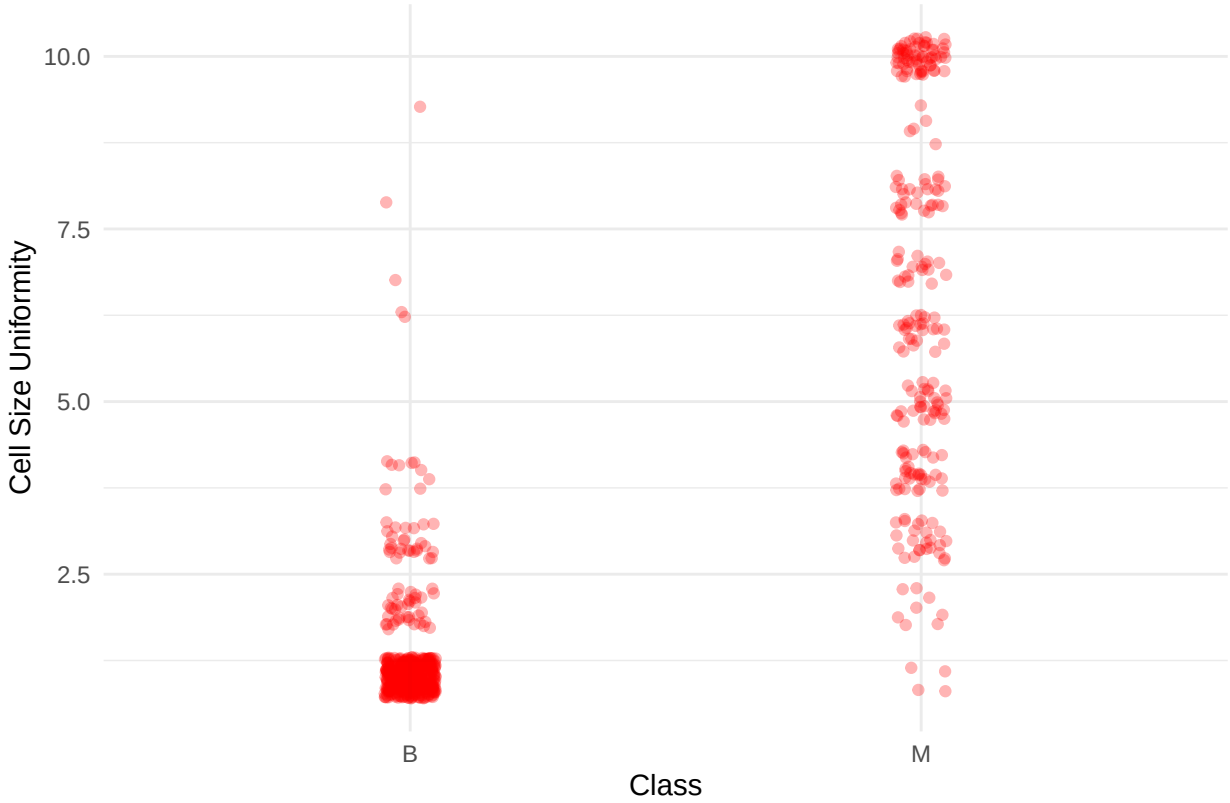

Figure 1: Cell size uniformity spread per class.

```
ggplot(data=plot_data, mapping = aes(x = Class, y = Cell_Shape_Uniformity)) +
  geom_jitter(width = 0.05, height = 0.3, col="green", alpha = 0.3) +
  labs(y="Cell Shape Uniformity",
       caption="Figure 2: Cell shape uniformity spread per class.") +
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

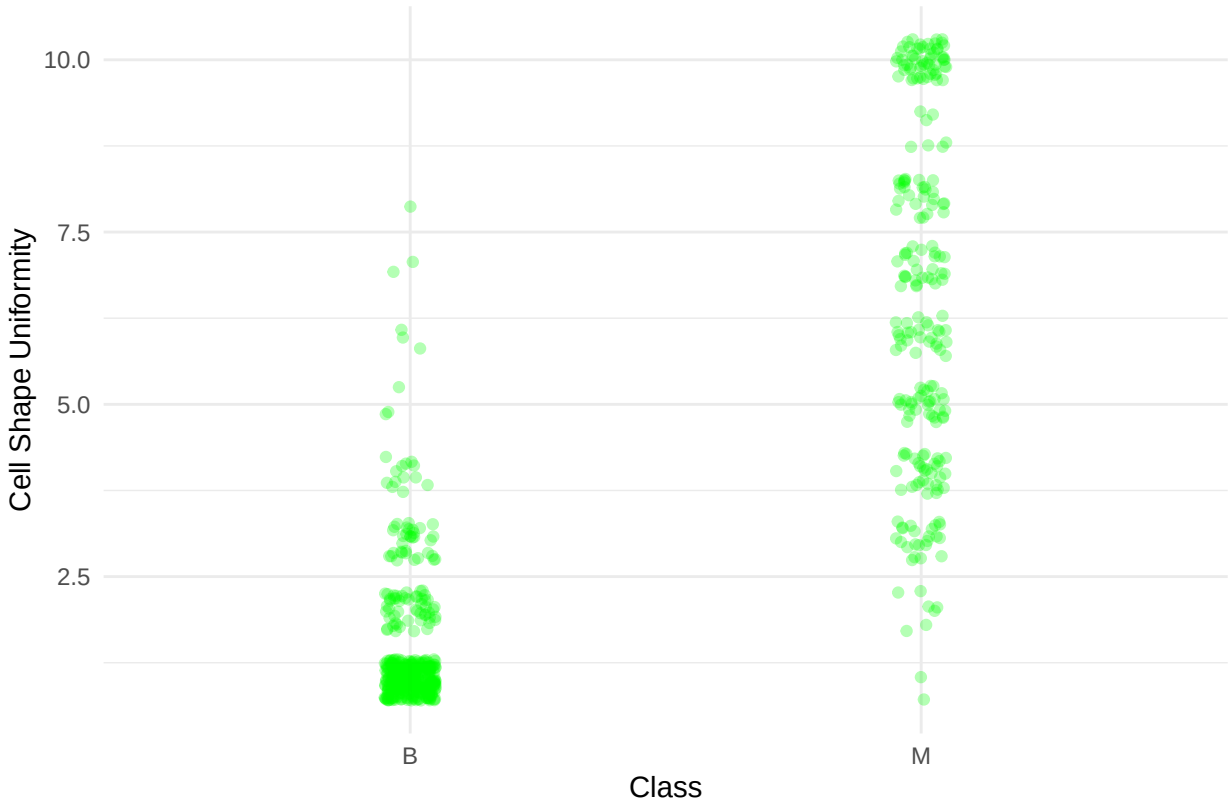

Figure 2: Cell shape uniformity spread per class.

```
ggplot(data=plot_data, mapping = aes(x = Class, y = Bare_Nuclei)) +
  geom_jitter(width = 0.05, height = 0.3, col="blue", alpha = 0.3) +
  labs(y="Nuclei With Cytoplasm",
       caption="Figure 3: Bare nuclei spread per class.") +
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

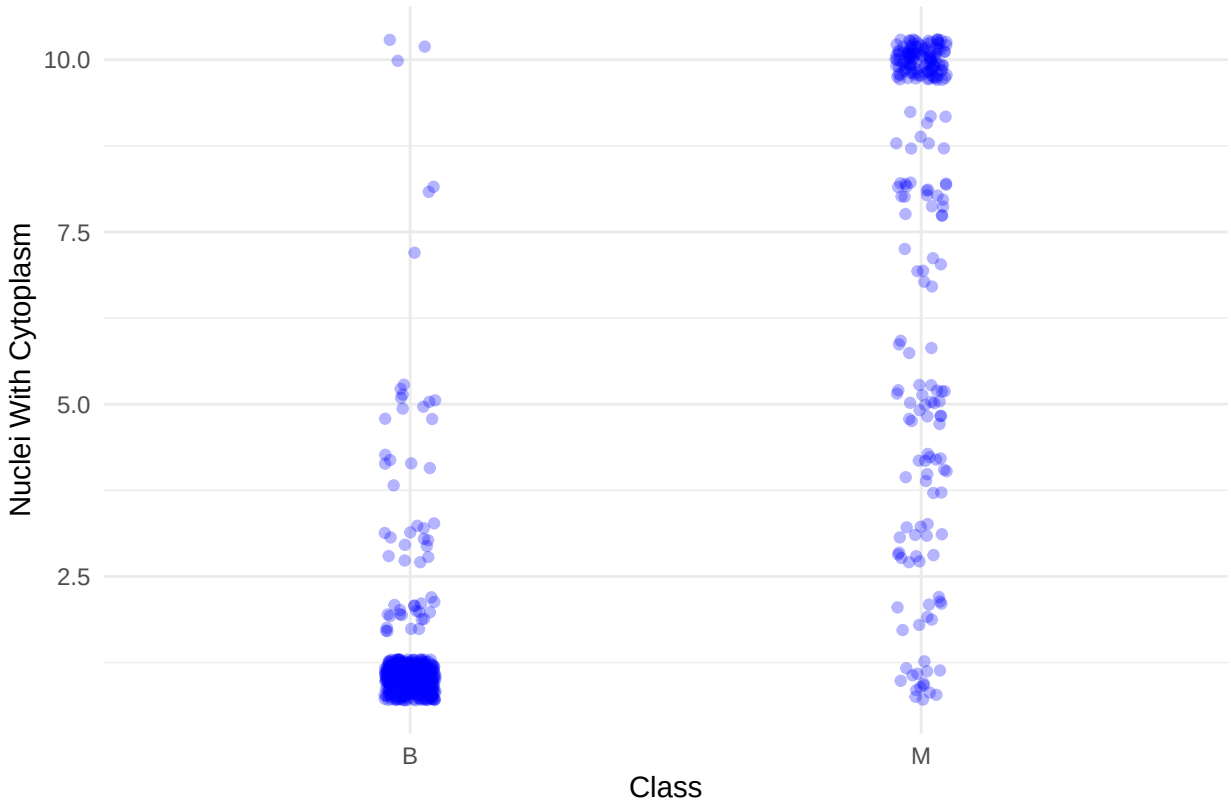

Figure 3: Bare nuclei spread per class.

It does seem to be the case that the more individual cells differ from each other (that is to say; being less uniform), the higher the chances are for the cancer to be malignant. The same goes for finding cytoplasm in the cell nucleus. However, it looks like these 3 attributes correlate quite strongly not only with the class, but also with each other. It would make sense for cell shape and size uniformity to covariate, but let us actually compare each of these features with each other.

```
ggplot(data=plot_data, mapping = aes(x = Cell_Size_Uniformity, y = Cell_Shape_Uniformity)) +
  geom_point(aes(col=Class)) +
  geom_smooth(method = "lm", se = T) +
  labs(y="Cell Shape Uniformity", x="Cell Size Uniformity",
       caption="Figure 4: Plot of the correlation between cell shape and size discrepancies") +
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

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

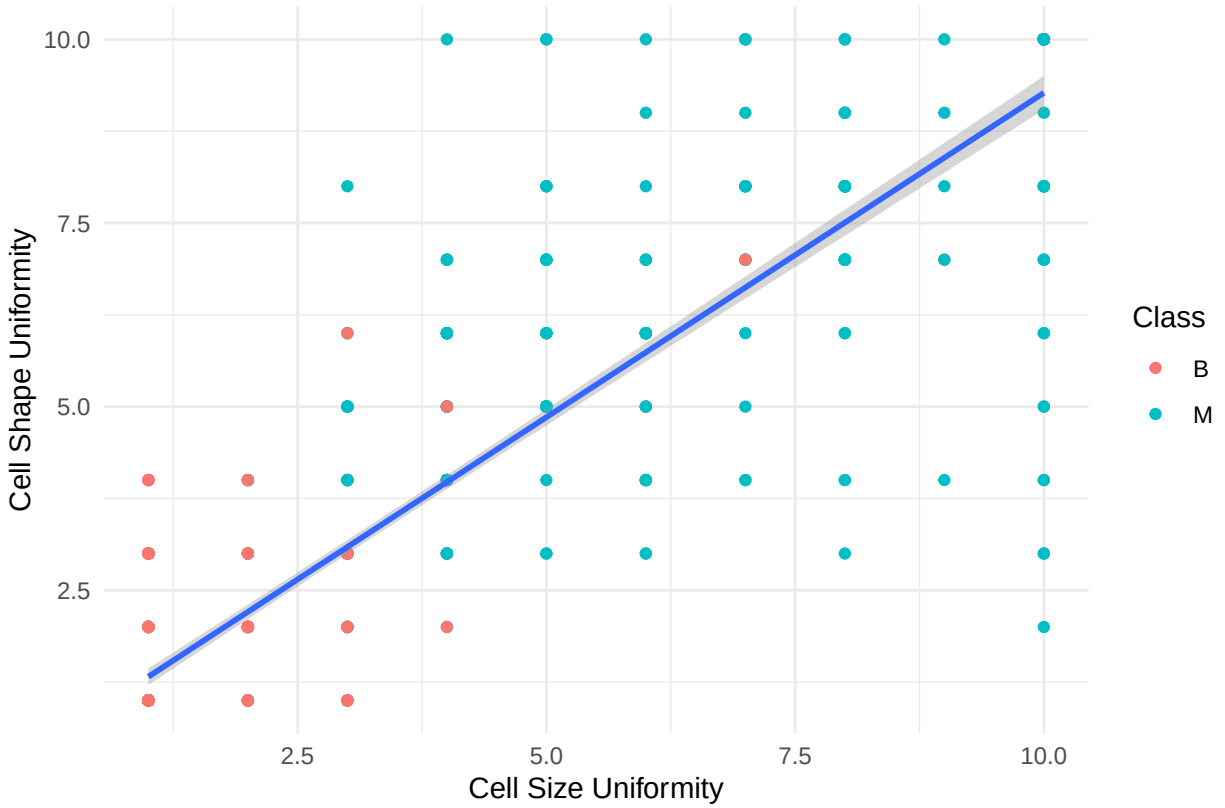

Figure 4: Plot of the correlation between cell shape and size discrepancies

Unsurprisingly, it does seem to be the case that the larger the differences in size between cells, the larger the differences in shapes as well. This makes sense as both these differences can be explained by erratic cell growth. Continuing to the other comparisons;

```
ggplot(data=plot_data, mapping = aes(x = Cell_Size_Uniformity, y = Bare_Nuclei)) +
  geom_point(aes(col=Class)) +
  geom_smooth(method = "lm", se = T) +
  labs(y="Nuclei With Cytoplasm", x="Cell Size Uniformity",
       caption="Figure 5: Plot of the correlation between nuclei with cytoplasm and cell size uniformity
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

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

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

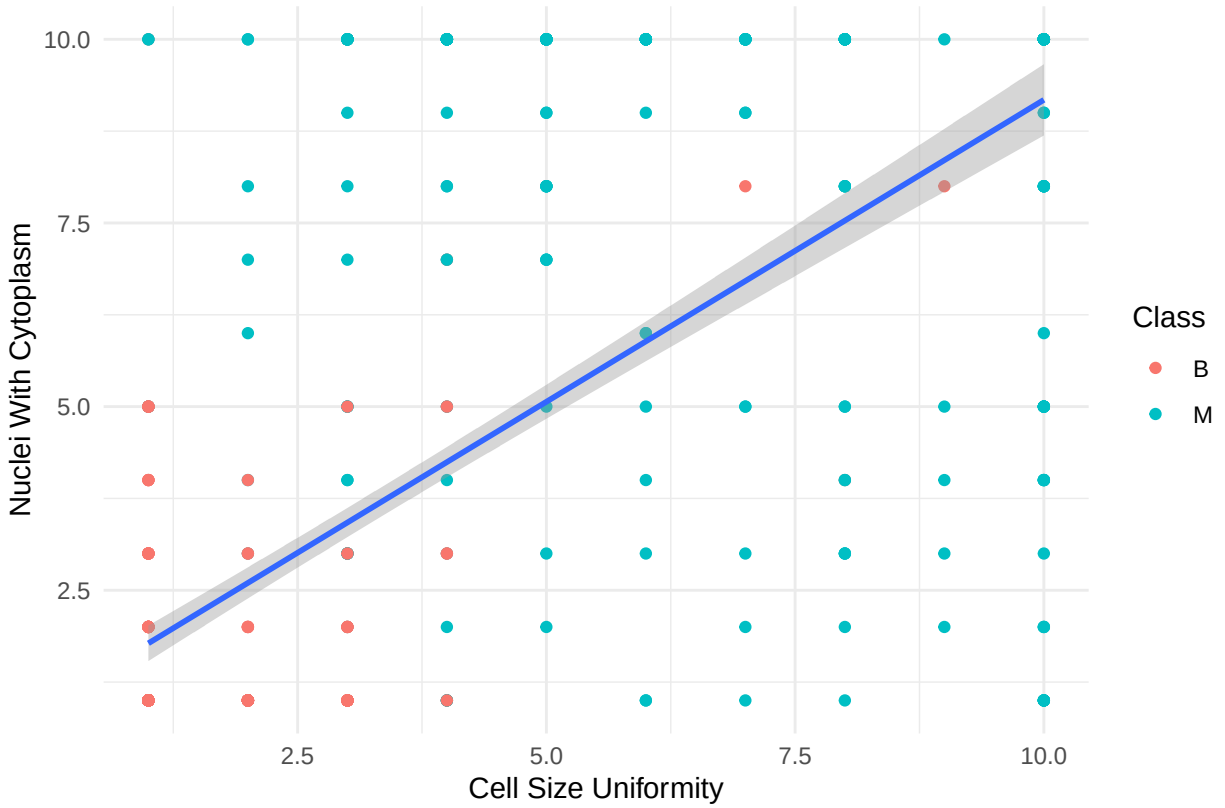

Figure 5: Plot of the correlation between nuclei with cytoplasm and cell size uniformity.

```
ggplot(data=plot_data, mapping = aes(x = Cell_Shape_Uniformity, y = Bare_Nuclei)) +
  geom_point(aes(col=Class)) +
  geom_smooth(method = "lm", se = T) +
  labs(y="Nuclei With Cytoplasm", x="Cell Shape Uniformity",
       caption="Figure 6: Plot of the correlation between nuclei with cytoplasm and cell shape uniformi
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

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

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

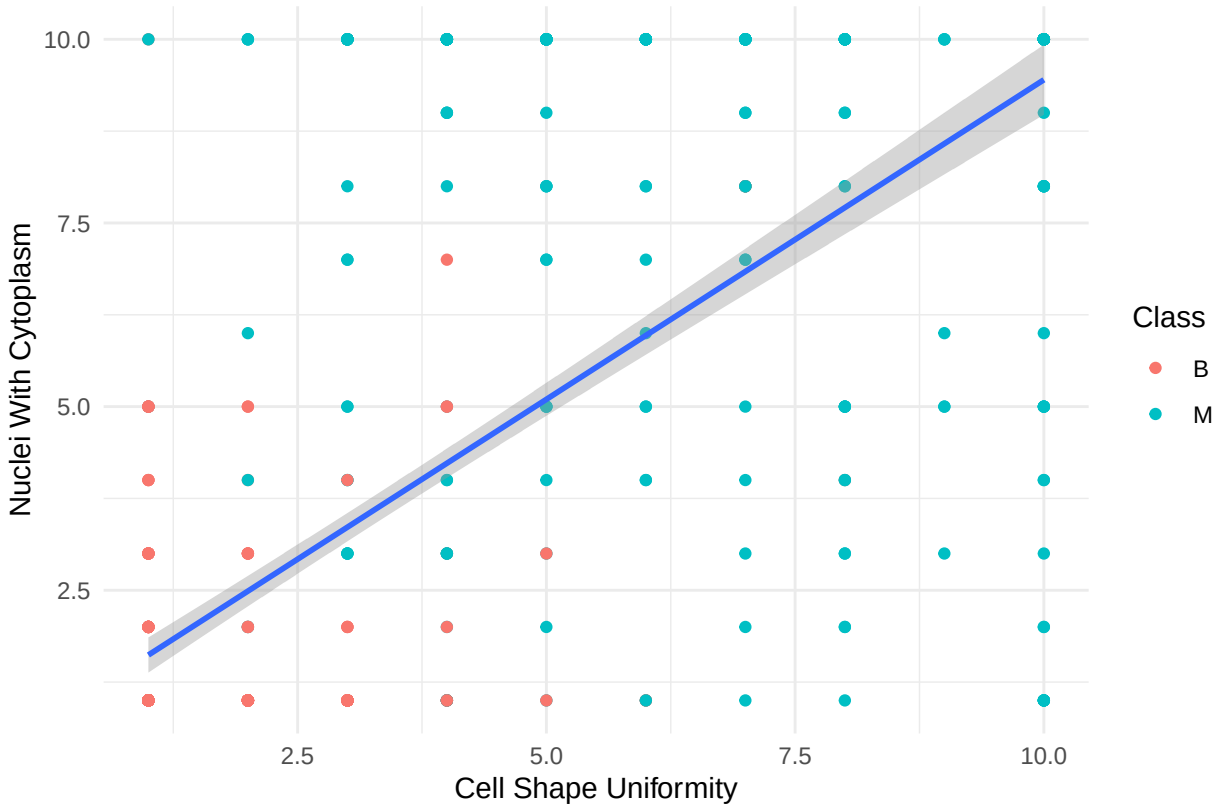

Figure 6: Plot of the correlation between nuclei with cytoplasm and cell shape uniformity.

The features that correlate less with the malignancy also seem to correlate slightly less amongst themselves, but let us make a heatmap of the entire dataset, to make sure we did not miss any other highly correlating attributes with our 80% cut-off.

```
cor_matrix <- round(cor(na.omit(data[,-1])),2)
melted_cor_matrix <- melt(cor_matrix)

ggplot(data = melted_cor_matrix, aes(x=Var1, y=Var2, fill=value, )) +
  geom_tile() +
  labs(y="", x="",
       caption="Figure 7: Heatmap showing the correlations between all attributes.") +
  scale_fill_gradient(low="black", high="red") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

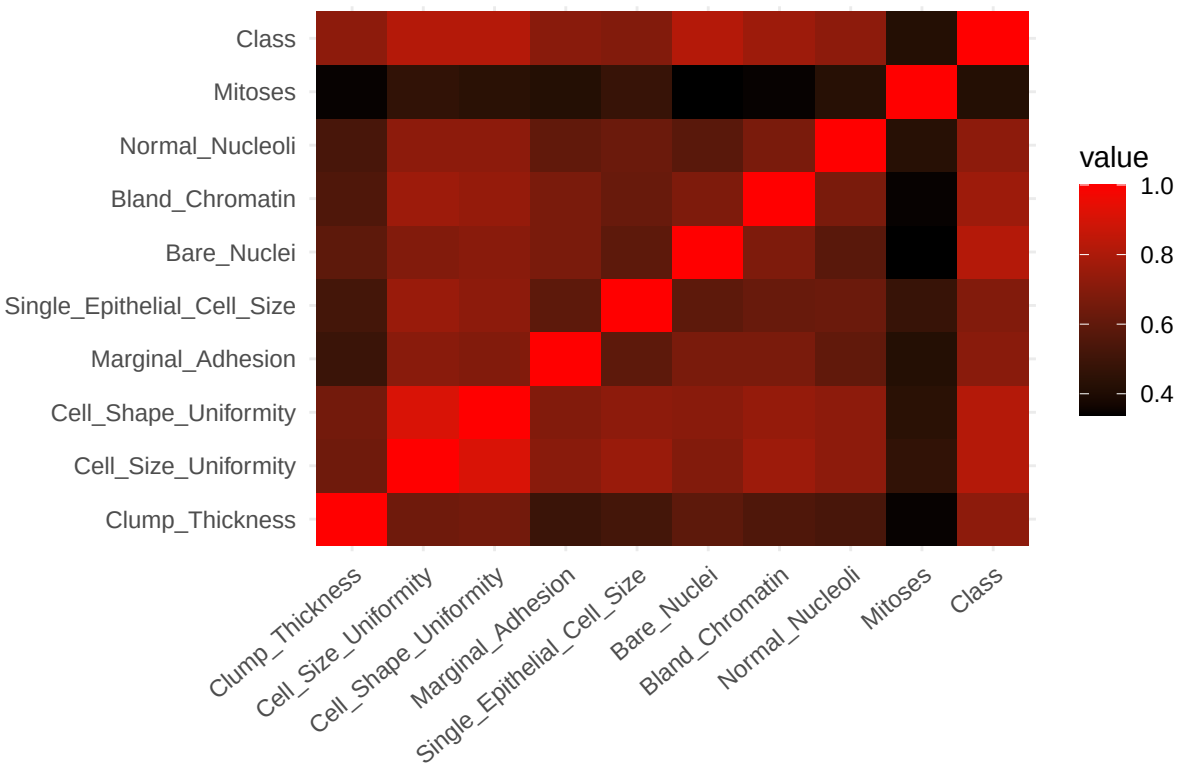

Figure 7: Heatmap showing the correlations between all attributes.

We can clearly see cell uniformity and bare nuclei at the top, but it looks like all features correlate at least somewhat, with bland chromatin seemingly at the top of the sub-80% attributes with a correlation to class of roughly 76%. The heatmap also confirms the correlation between cell size and shape uniformity.

Having learnt this, let us now also make a graph of the spread for bland chromatin.

```
ggplot(data=plot_data, mapping = aes(x = Class, y = Bland_Chromatin)) +
  geom_jitter(width = 0.05, height = 0.3, col="cyan", alpha = 0.3) +
  labs(y="Bland Chromatin",
       caption="Figure 8: Bland Chromatin spread per class.") +
  theme_minimal() +
  theme(plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0.5))
```

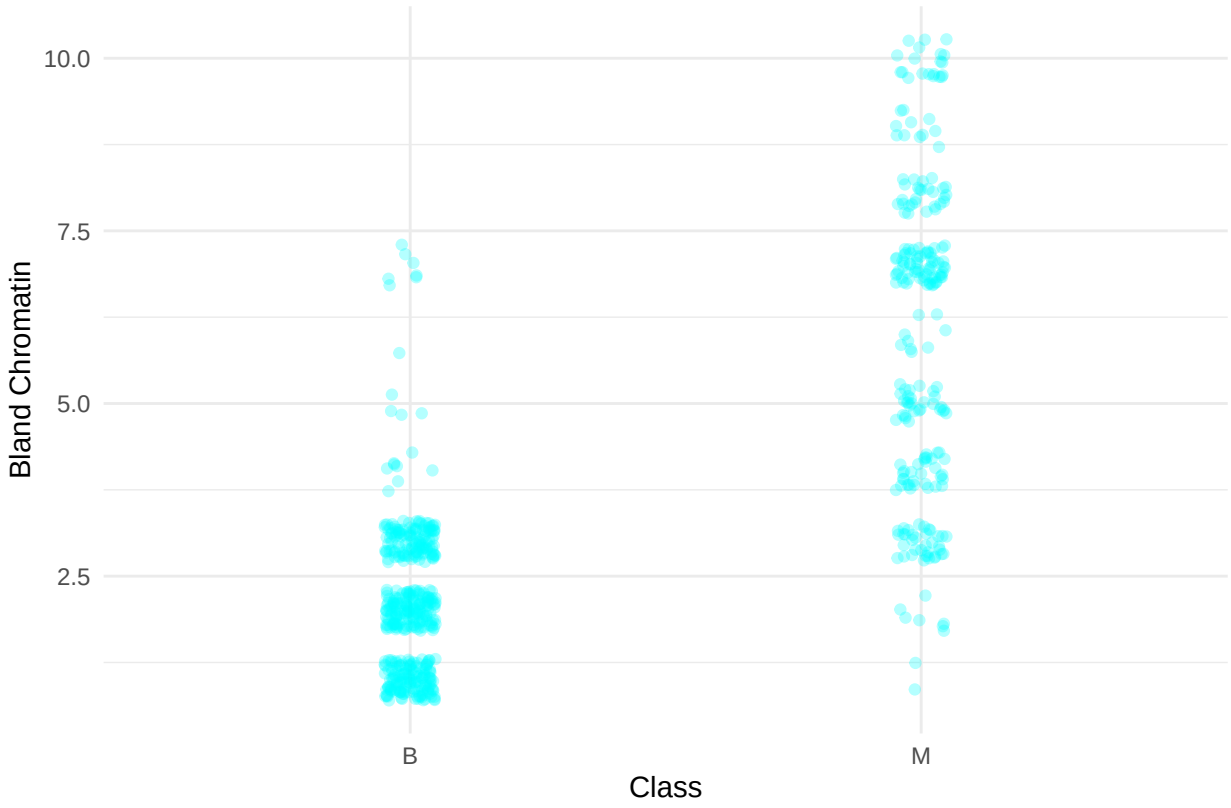

Figure 8: Bland Chromatin spread per class.

This seems to somewhat mimic the spreads of the top 3 correlating features, albeit to a slightly lesser degree. Take note that a grade of 10 on the bland chromatin attribute means that the chromatin is maximally coarse and a 1 stands for very finely textured chromatin.

There seem to be some duplicate entries in the data set, to remedy this I shall only take the unique rows.

```
clean_data <- unique(plot_data)
```

It is important to get rid of duplicates like these, because they increase the weight of their corresponding attribute values.

As could be seen in the heatmap in figure 7, mitosis has by far the lowest correlation with malignancy and will therefore be removed from the data set. Aside from this, it is also unclear what "abnormal" mitosis means exactly. As for the normal nucleoli attribute, its precise definition is unclear. A grade of 1, which is to say a "normal" nucleolus, is defined as being small and barely visible. There is zero clarification, however, of how a nucleolus starts to differ when looking at the higher grades. One can assume that there might then be multiple nucleoli, or they might be larger, but this is never quantified. I deem this too unclear and shall also remove this column from the data set.

```
clean_data$Mitoses <- NULL
clean_data$Normal_Nucleoli <- NULL
```