Thema 9 Log: Breast Cancer Wisconsin (Original) Data Set

Naomi Hindriks

9/21/2021

EDA : Breast Cancer Wisconsin (Original) Data Set

Data description

The data set: **Breast Cancer Wisconsin (Original) Data Set** is downloaded from the UCI machine learning repository. The data were collected by the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg.

The UCI website states that the data set contains 699 instances. According to the corresponding *breast-cancer-wisconsin.names* file (also downloaded from the UCI website) each instance is made up of 10 attributes, plus the class attribute. More detailed information of these attributes is shown in Table 1. The information found in Table 1 is a combination of information that was found in the *breast-cancer-wisconsin.names* file and User Manual Breast Cancer Diagnosis Web User Interface that includes an explanation on how to score the cytological characteristic.

According to the *breast-cancer-wisconsin.names* file there are 16 instances that contain a single missing attribute value, these are represented by "?" characters in the data file. It also states that out of the 699 instances there are 458 (65.5%) classified as benign and 241 (34.5%) classified as malignant.

To ensure the continued availability of the data and names files, they were copied to a personal repository.

```
attribute.info <- read.csv("attribute_info.csv", sep=";")

attribute.info.temp <- data.frame(
  "Column" = attribute.info$column,
  "Attribute" = attribute.info$full.name,
  "Unit" = attribute.info$unit,
  "Description" = attribute.info$description
  )

kbl(
    attribute.info.temp,
    row.names = F,
    caption = "Attribute Information. The cytological characteristics of breast FNAs (seen in rows 2-10)
    booktabs = T,
    linesep = "",
    longtable = T
  ) %>%
  kable_styling(latex_options = c("striped")) %>%
  column_spec(1:3, width = "1.5cm") %>%
  column_spec(4, width = "10cm")
```

Table 1: Attribute Information. The cytological characteristics of breast FNAs (seen in rows 2-10) get a score from 1 to 10 by an examining physician with 1 being the closest to benign and 10 the most anaplastic.

Column	Attribute	Unit	Description
1	Sample code number	id number	Unique number given to each sample
2	Clump Thickness	1-10	Assesses if cells are mono or multi-layered
3	Uniformity of Cell Size	1-10	Evaluate the consistency in size of the cells in the sample
4	Uniformity of Cell Shape	1-10	Evaluate the consistency in shape of the cells in the sample
5	Marginal Adhesion	1-10	Quantifies proportion of cells that stick together
6	Single Epithelial Cell Size	1-10	Measures the enlargement of epithelial cells size
7	Bare Nuclei	1-10	Proportion of nuclei surrounded by cytoplasm versus those that are not
8	Bland Chromatin	1-10	Rates the uniform texture of the nucleus in a range from fine to coarse
9	Normal Nucleoli	1-10	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful
10	Mitoses	1-10	Describes the level of mitotic activity
11	Class	2 or 4	Classification: 2 for benign and 4 for malignant

```
#clean up environment
remove(attribute.info.temp)
```

Data loading and prepping

```
data <- read.table(file = 'data/breast-cancer-wisconsin.data',
                                     header = F,
                                     sep = ",",
                                     na.strings = '?')

str(data)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ V1 : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ...
##  $ V2 : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V3 : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V4 : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V5 : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V6 : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V7 : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V8 : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V9 : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V10: int  1 1 1 1 1 1 1 1 5 1 ...
##  $ V11: int  2 2 2 2 2 4 2 2 2 2 ...
```

The data has been loaded, but it can be seen that the column names were not included in the data file. Furthermore the data does not seem to be of the correct data type, columns 2-10 should all be (ordered) factors. To give the columns the correct names and have easy access to the column descriptions I have created a simple csv file (attribute_info.csv).

```r
names(data) <- attribute.info$name

data$class <- factor(data$class, levels = c(2, 4), labels = c("Benign", "Malignant"))

for(col.name in names(data)[2:10]) {
  data[, col.name] <- factor(data[, col.name], levels=1:10, ordered=T)
}

str(data)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ id                    : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 10330
##  $ clump.thick           : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 3 6 4 8 1 2 2 4 ...
##  $ uni.cell.size         : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 1 1 2 ...
##  $ uni.cell.shape        : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 2 1 1 ...
##  $ marg.adhesion         : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 5 1 1 3 8 1 1 1 1 ...
##  $ single.epith.cell.size: Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2 7 2 3 2 7 2 2 2 2 ...
##  $ bare.nuclei           : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 10 2 4 1 10 10 1 1 1 ...
##  $ bland.chrom           : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 3 3 3 3 3 9 3 3 1 2 ...
##  $ norm.nucleoli         : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses               : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 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
#clean up environment
remove(col.name)
```

Now the columns have names and the values are of the correct data type.

Data verification

The original data description stated that 699 instances with 10 attributes + a class label are present.

```r
dim(data)
```

```
## [1] 699  11
```

This checks out. The original data description also stated that there are 16 instances with a single missing value, the instances that have a mising value will be removed from the data set. This means there are no more than 16 missing values and the number of complete cases should be 699 - 16.

```
sum(is.na(data))
```

```
## [1] 16
```

```
complete.instances <- complete.cases(data)

699 - sum(complete.instances)
```

```
## [1] 16
```

This is correct. According to the original data description the class distribution is as follows: benign: 458 (65.5%), malignant: 241 (34.5%).

```
summary(data$class)
```

```
##    Benign Malignant
##       458       241
```

```
format(summary(data$class) / nrow(data) * 100, digits = 3)
```

```
##    Benign Malignant
##    "65.5"    "34.5"
```

Again this checks out. The last thing that I am going to check are the *Sample code numbers* of instances. The data original description stated that this is an id number, therefore I assume all of these numbers should be unique.

```
length(unique(data$id))
```

```
## [1] 645
```

The number of unique *sample code numbers* is 645, which is less than the 699 instances in the data. This is odd and requires further investigation. According to the original data description the data set is divided in 8 different groups, each group being collected in a different period of time. The groups 1 to 8 contain 367, 70, 31, 17, 48, 49, 31 and 86 instances respectively. Perhaps the *sample code numbers* of the instances are unique within their group.

```
group.sizes <- c(367, 70, 31, 17, 48, 49, 31, 86)
duplicates.per.group <- c()

current.slice.start <- 0
i <- 0

for (group.size in group.sizes) {
  i <- i + 1
  group.row.numbers <- (current.slice.start + 1):(current.slice.start + group.size)
  current.slice.start <- current.slice.start + group.size

  duplicates <- sum(duplicated(data$id[group.row.numbers]))
  duplicates.per.group <- c(duplicates.per.group, duplicates)

  print(paste("Group ", i, ": ", duplicates , sep = ""))
}
```

```
## [1] "Group 1: 20"
## [1] "Group 2: 5"
## [1] "Group 3: 1"
## [1] "Group 4: 0"
## [1] "Group 5: 2"
## [1] "Group 6: 2"
## [1] "Group 7: 0"
## [1] "Group 8: 6"
```

```
# The total of duplicates when only looking inside group
sum(duplicates.per.group)
```

```
## [1] 36
```

```
# The total duplicates in and outside group
nrow(data) - length(unique(data$id))
```

```
## [1] 54
```

```
#clean up environment
remove(group.sizes, duplicates.per.group,
       current.slice.start, i, group.size,
       group.row.numbers, duplicates)
```

When looking at these numbers it is clear that there are duplicates within the groups and duplicates between different groups. This means that it is not logical that the duplicates are just duplicated rows that somehow got copied an extra time, because if that were the case we would expect to see only duplicates within groups. It is also not logical that the *sample code numbers* are reused in different groups since there are also duplicates within the groups. The next step is to check if the instances with duplicated *sample code numbers* have every attribute duplicated.

```
nrow(data[duplicated(data), ])
```

```
## [1] 8
```

There are 8 rows that are an exact copy of another row, this means that there are instances with the same *sample code number* but different values for the other attributes. Tables 2-47 show the instances that share their *sample code number* with at least one other instance. The tables that have duplicates where every attribute is the same have a red header.

```
duplicated.ids <- unique(data$id[duplicated(data$id)])

for (duplicate.id in duplicated.ids) {
  duplicate.entries.temp <- data[data$id == duplicate.id, ]

  if (sum(duplicated(duplicate.entries.temp)) > 0) {
    header.color <- "red"
  } else {
    header.color <- "white"
  }
```

```r
  table <- kbl(
    duplicate.entries.temp,
    row.names = T,
    col.names = attribute.info$full.name,
    caption = paste("Instances with duplicate id:", duplicate.id),
    booktabs = T,
    linesep = ""
  ) %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position")) %>%
  column_spec(1:11, width = "1.5cm") %>%
  row_spec(0, background = header.color)

  print(table)
}
```

Table 2: Instances with duplicate id: 1033078

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
9	1033078	2	1	1	1	2	1	1	1	5	Benign
10	1033078	4	2	1	1	2	1	2	1	1	Benign

Table 3: Instances with duplicate id: 1070935

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
30	1070935	1	1	3	1	2	1	1	1	1	Benign
31	1070935	3	1	1	1	1	1	2	1	1	Benign

Table 4: Instances with duplicate id: 1143978

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
82	1143978	4	1	1	2	2	1	2	1	1	Benign
83	1143978	5	2	1	1	2	1	3	1	1	Benign

Table 5: Instances with duplicate id: 1171710

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
109	1171710	1	1	1	1	2	1	2	3	1	Benign
110	1171710	6	5	4	4	3	9	7	8	3	Malignant

Table 6: Instances with duplicate id: 1173347

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
116	1173347	1	1	1	1	2	5	1	1	1	Benign
117	1173347	8	3	3	1	2	2	3	2	1	Benign

Table 7: Instances with duplicate id: 1174057

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
121	1174057	1	1	2	2	2	1	3	1	1	Benign
122	1174057	4	2	1	1	2	2	3	1	1	Benign

Table 8: Instances with duplicate id: 1212422

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
195	1212422	3	1	1	1	2	1	3	1	1	Benign
196	1212422	4	1	1	1	2	1	3	1	1	Benign

Table 9: Instances with duplicate id: 1218860

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
208	1218860	1	1	1	1	1	1	3	1	1	Benign
209	1218860	1	1	1	1	1	1	3	1	1	Benign

Table 10: Instances with duplicate id: 1017023

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
5	1017023	4	1	1	3	2	1	3	1	1	Benign
253	1017023	6	3	3	5	3	10	3	5	3	Benign

Table 11: Instances with duplicate id: 1100524

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
43	1100524	6	10	10	2	8	10	7	3	3	Malignant
254	1100524	6	10	10	2	8	10	7	3	3	Malignant

Table 12: Instances with duplicate id: 1116116

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
63	1116116	9	10	10	1	10	8	3	3	1	Malignant
255	1116116	9	10	10	1	10	8	3	3	1	Malignant

Table 13: Instances with duplicate id: 1168736

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
105	1168736	10	10	10	10	10	1	8	8	8	Malignant
256	1168736	5	6	6	2	4	10	3	6	1	Malignant

Table 14: Instances with duplicate id: 1182404

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
137	1182404	4	1	1	1	2	1	2	1	1	Benign
257	1182404	3	1	1	1	2	1	1	1	1	Benign
258	1182404	3	1	1	1	2	1	2	1	1	Benign
266	1182404	5	1	4	1	2	1	3	2	1	Benign
449	1182404	1	1	1	1	1	1	1	1	1	Benign
498	1182404	4	2	1	1	2	1	1	1	1	Benign

Table 15: Instances with duplicate id: 1198641

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
169	1198641	3	1	1	1	2	1	3	1	1	Benign
259	1198641	3	1	1	1	2	1	3	1	1	Benign
267	1198641	10	10	6	3	3	10	4	3	2	Malignant

Table 16: Instances with duplicate id: 320675

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
268	320675	3	3	5	2	3	10	7	1	1	Malignant
273	320675	3	3	5	2	3	10	7	1	1	Malignant

Table 17: Instances with duplicate id: 733639

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
322	733639	3	1	1	1	2	NA	3	1	1	Benign
323	733639	3	1	1	1	2	1	3	1	1	Benign

Table 18: Instances with duplicate id: 704097

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
315	704097	1	1	1	1	1	1	2	1	1	Benign
339	704097	1	1	1	1	1	1	2	1	1	Benign

Table 19: Instances with duplicate id: 493452

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
372	493452	1	1	3	1	2	1	1	1	1	Benign
373	493452	4	1	2	1	2	1	2	1	1	Benign

Table 20: Instances with duplicate id: 560680

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
291	560680	1	1	1	1	2	1	1	1	1	Benign
375	560680	3	1	2	1	2	1	2	1	1	Benign

Table 21: Instances with duplicate id: 1114570

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
388	1114570	5	3	3	2	3	1	3	1	1	Benign
389	1114570	2	1	1	1	2	1	2	2	1	Benign

Table 22: Instances with duplicate id: 1158247

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
94	1158247	1	1	1	1	2	1	2	1	1	Benign
394	1158247	1	1	1	1	1	1	1	1	1	Benign

Table 23: Instances with duplicate id: 1276091

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
242	1276091	3	1	1	3	1	1	3	1	1	Benign
430	1276091	2	1	1	1	2	1	2	1	1	Benign
431	1276091	1	3	1	1	2	1	2	2	1	Benign
432	1276091	5	1	1	3	4	1	3	2	1	Benign
463	1276091	6	1	1	3	2	1	1	1	1	Benign

Table 24: Instances with duplicate id: 1293439

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
434	1293439	3	2	2	3	2	1	1	1	1	Benign
435	1293439	6	9	7	5	5	8	4	2	1	Benign

Table 25: Instances with duplicate id: 734111

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
443	734111	1	1	1	3	2	3	1	1	1	Benign
444	734111	1	1	1	1	2	2	1	1	1	Benign

Table 26: Instances with duplicate id: 1105524

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
48	1105524	1	1	1	1	2	1	2	1	1	Benign
469	1105524	4	1	1	1	2	1	1	1	1	Benign

Table 27: Instances with duplicate id: 1115293

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
62	1115293	1	1	1	1	2	2	2	1	1	Benign
491	1115293	1	1	1	1	2	1	1	1	1	Benign

Table 28: Instances with duplicate id: 1320077

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
517	1320077	1	1	1	1	1	1	1	1	1	Benign
518	1320077	1	1	1	1	1	1	2	1	1	Benign

Table 29: Instances with duplicate id: 769612

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
526	769612	3	1	1	2	2	1	1	1	1	Benign
527	769612	4	1	1	1	2	1	1	1	1	Benign

Table 30: Instances with duplicate id: 798429

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
338	798429	1	1	1	1	2	1	3	1	1	Benign
528	798429	4	1	1	1	2	1	3	1	1	Benign

Table 31: Instances with duplicate id: 1116192

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
65	1116192	1	1	1	1	2	1	2	1	1	Benign
538	1116192	5	1	2	1	2	1	3	1	1	Benign

Table 32: Instances with duplicate id: 1240603

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
548	1240603	2	1	1	1	1	1	1	1	1	Benign
549	1240603	3	1	1	1	1	1	1	1	1	Benign

Table 33: Instances with duplicate id: 1299924

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
512	1299924	5	1	1	1	2	1	2	1	1	Benign
553	1299924	3	2	2	2	2	1	4	2	1	Benign

Table 34: Instances with duplicate id: 1321942

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
561	1321942	5	1	1	1	2	1	3	1	1	Benign
562	1321942	5	1	1	1	2	1	3	1	1	Benign

Table 35: Instances with duplicate id: 385103

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
270	385103	1	1	1	1	2	1	3	1	1	Benign
576	385103	5	1	2	1	2	1	3	1	1	Benign

Table 36: Instances with duplicate id: 411453

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
272	411453	5	1	1	1	2	1	3	1	1	Benign
608	411453	1	1	1	1	2	1	1	1	1	Benign

Table 37: Instances with duplicate id: 822829

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
345	822829	7	6	4	8	10	10	9	5	3	Malignant
613	822829	8	10	10	10	6	10	10	10	10	Malignant

Table 38: Instances with duplicate id: 1061990

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
536	1061990	1	1	3	2	2	1	3	1	1	Benign
619	1061990	4	1	1	1	2	1	2	1	1	Benign

Table 39: Instances with duplicate id: 1238777

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
472	1238777	6	1	1	3	2	1	1	1	1	Benign
633	1238777	1	1	1	1	2	1	1	1	1	Benign

Table 40: Instances with duplicate id: 1277792

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
639	1277792	4	1	1	1	2	1	1	1	1	Benign
640	1277792	5	1	1	3	2	1	1	1	1	Benign

Table 41: Instances with duplicate id: 1299596

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
468	1299596	6	6	6	5	4	10	7	6	2	Malignant
645	1299596	2	1	1	1	2	1	1	1	1	Benign

Table 42: Instances with duplicate id: 1339781

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
661	1339781	1	1	1	1	2	1	2	1	1	Benign
662	1339781	4	1	1	1	2	1	3	1	1	Benign

Table 43: Instances with duplicate id: 1354840

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
673	1354840	2	1	1	1	2	1	3	1	1	Benign
674	1354840	5	3	2	1	3	1	1	1	1	Benign

Table 44: Instances with duplicate id: 466906

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
684	466906	1	1	1	1	2	1	1	1	1	Benign
685	466906	1	1	1	1	2	1	1	1	1	Benign

Table 45: Instances with duplicate id: 654546

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
690	654546	1	1	1	1	2	1	1	1	8	Benign
691	654546	1	1	1	3	2	1	1	1	1	Benign

Table 46: Instances with duplicate id: 695091

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
578	695091	1	1	1	1	2	1	2	1	1	Benign
692	695091	5	10	10	5	4	5	4	4	1	Malignant

Table 47: Instances with duplicate id: 897471

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
698	897471	4	8	6	4	3	4	10	6	1	Malignant
699	897471	4	8	8	5	4	5	10	4	1	Malignant

```r
#clean up environment
remove(duplicate.entries.temp, table, header.color, duplicated.ids, duplicate.id)
```

When inspecting these tables it becomes apparent that the duplicated data is sometimes in consecutive rows, but not always. It can also be observed that most of the instances with duplicated *sample code numbers* have the same class label, but not always. You can also see that most duplicates come in pairs, but they also come in bigger groups, up to 6 instances in one group (see Table 14). I do not see any pattern in how these rows are duplicated, nor can I think of any logical explanation for this. Since I do not want to risk using instances that are from the same person or sample I will keep only one instance per *sample code number*, and remove the duplicated rows.

Removing data

First the instances with a missing value will be removed. After that the instances with a duplicated *sample code number* will be removed.

```
# Keep unfiltered data in variable
unfiltered.data <- data

# only keep rows with complete instances
data <- data[complete.instances, ]

# verify 16 instances have been removed
dim(data)
```

```
## [1] 683  11
```

```
#clean up environment
remove(complete.instances)
```

After removing the instances with a missing value, there are 683 instances left, which is as expected because 699 - 16 = 683

```
# find instances with duplicated id
duplicates <- duplicated(data$id)

# remove duplicate instances from data
data <- data[!duplicates, ]

# Making the id the rowname and removing id column
row.names(data) <- data$id
data <- data[2:11]

# print what the dimensions are after cleaning the data
dim(data)
```

```
## [1] 630  10
```

Then after removing the instances with duplicated *sample code numbers* there are 630 instances left, these instances do not have missing values or duplicated *sample code numbers.*

Exploring variables

A first scan of the attributes.

```
summary(data)
```

```
##   clump.thick  uni.cell.size uni.cell.shape marg.adhesion
##  1      :127   1      :339   1      :312    1      :355
##  5      :118   10     : 62   10     : 54    2      : 54
##  3      : 96   3      : 47   2      : 52    10     : 54
##  4      : 68   2      : 40   3      : 51    3      : 53
##  10     : 68   4      : 38   4      : 41    4      : 32
##  2      : 47   5      : 29   5      : 31    8      : 25
##  (Other):106   (Other): 75   (Other): 89    (Other): 57
##  single.epith.cell.size  bare.nuclei   bland.chrom  norm.nucleoli    mitoses
##  2      :343             1      :363   2      :149   1      :395   1      :515
##  3      : 66             10     :126   3      :145   10     : 59   2      : 34
##  4      : 44             3      : 28   1      :133   3      : 39   3      : 30
##  6      : 39             5      : 28   7      : 68   2      : 29   10     : 13
##  5      : 38             2      : 27   4      : 35   8      : 22   4      : 12
##  1      : 37             4      : 19   5      : 34   6      : 21   7      :  9
##  (Other): 63             (Other): 39   (Other): 66   (Other): 65   (Other): 17
##        class
##  Benign   :400
##  Malignant:230
##
##
##
##
##
```

For all attributes (except the class attribute) the most common value is 1 or 2. This makes sense, the most instances are classified as benign and lower numbers indicate more benign characteristics. Now I will make a table and bargraph to compare the class distribution before and after the filtering.

```
class.distribution <- data.frame(
  filtered = c(rep("before", nrow(unfiltered.data)), rep("after", nrow(data))),
  class = c(as.character(unfiltered.data$class), as.character(data$class)))


d1 <-  class.distribution %>% group_by(filtered, class) %>%
  tally %>%
  bind_rows(class.distribution %>% group_by(filtered) %>%
      tally %>%
      mutate(class="Total")) %>%
  mutate(pct = round(n/((sum(n)/2))*100)) %>%
  arrange(desc(filtered))

kbl(
  d1[,2:4],
  caption = "Data distribution before and after filtering, n is the number of instances and pct is the
  kable_styling(latex_options = c("HOLD_position")) %>%
  pack_rows(index = table(fct_inorder(d1$filtered)))
```

Table 48: Data distribution before and after filtering, n is the number of instances and pct is the percentage of instances.

class	n	pct
before		
Benign	458	66
Malignant	241	34
Total	699	100
after		
Benign	400	63
Malignant	230	37
Total	630	100

```r
ggplot(class.distribution, aes_string(x = "class", y = "..prop..")) +
  geom_bar(
    aes(fill = factor(filtered), group = -as.numeric(factor(filtered))),
      position = position_dodge()
      ) +
  geom_text(
    aes(label = ..count.., group = -as.numeric(factor(filtered))),
    stat = "count",
    position = position_dodge(width = 0.9),
    vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  scale_fill_manual(
      name = "Data set",
      values = c(hue_pal()(2)),
      breaks = c("before", "after"),
      labels = c("Data before filtering", "Data after filtering")) +
  labs(title="Class distribution before and after filtering of the data set") +
  xlab("Class") +
  ylab("Pecentage of data set")
```

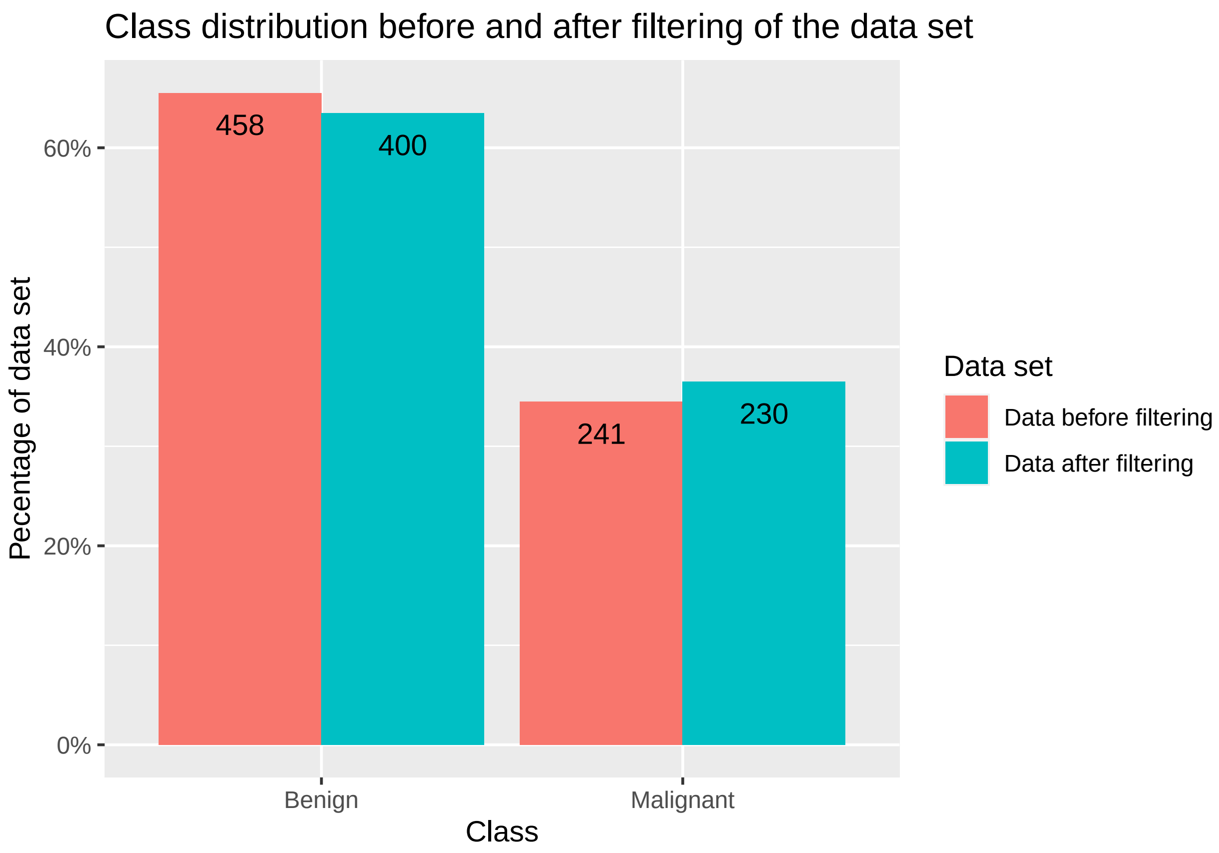

Figure 1: Distribution of the class attribute of the data before and after the filtering steps (the filtering steps are: removing rows with missing data, and than removing duplicated sample code numbers). The numbers in the bars are the actual number of instances in the data set.

When looking at *Table 48* and *Figure 1* you can see that the class distribution hasn't changed all that much from the filtering steps. This is a positive sign, if it had changed a lot the distribution after filtering might not have been representative anymore. Luckily that is not the case.

I will make bar plots to show the distribution of the cytological characteristic. I chose bar plots for this because the data is ordinal.

```
long.data <- pivot_longer(data, 1:9)

names.labs <- attribute.info$full.name
names(names.labs) <- attribute.info$name

ggplot(long.data, aes(x=value)) +
  geom_bar(aes(y = ..prop.., fill = name, group = class), stat="count") +
  labs(y = "Percent", fill="Attribute") +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_discrete(
    name = "Attribute",
    breaks = sort(attribute.info$name),
    labels = attribute.info$full.name[order(attribute.info$name)]
    ) +
```

```r
  labs(
    title="Distribution of cytological characteristics scores",
    x ="Score on a scale from 1 to 10",
    y = "Percentage of instances"
    ) +
  facet_grid(
    name ~ class,
    scales = "free",
    margin = "class",
    labeller = labeller(name = names.labs)
  ) +
  theme(legend.position = "bottom", strip.text.y = element_blank())
```

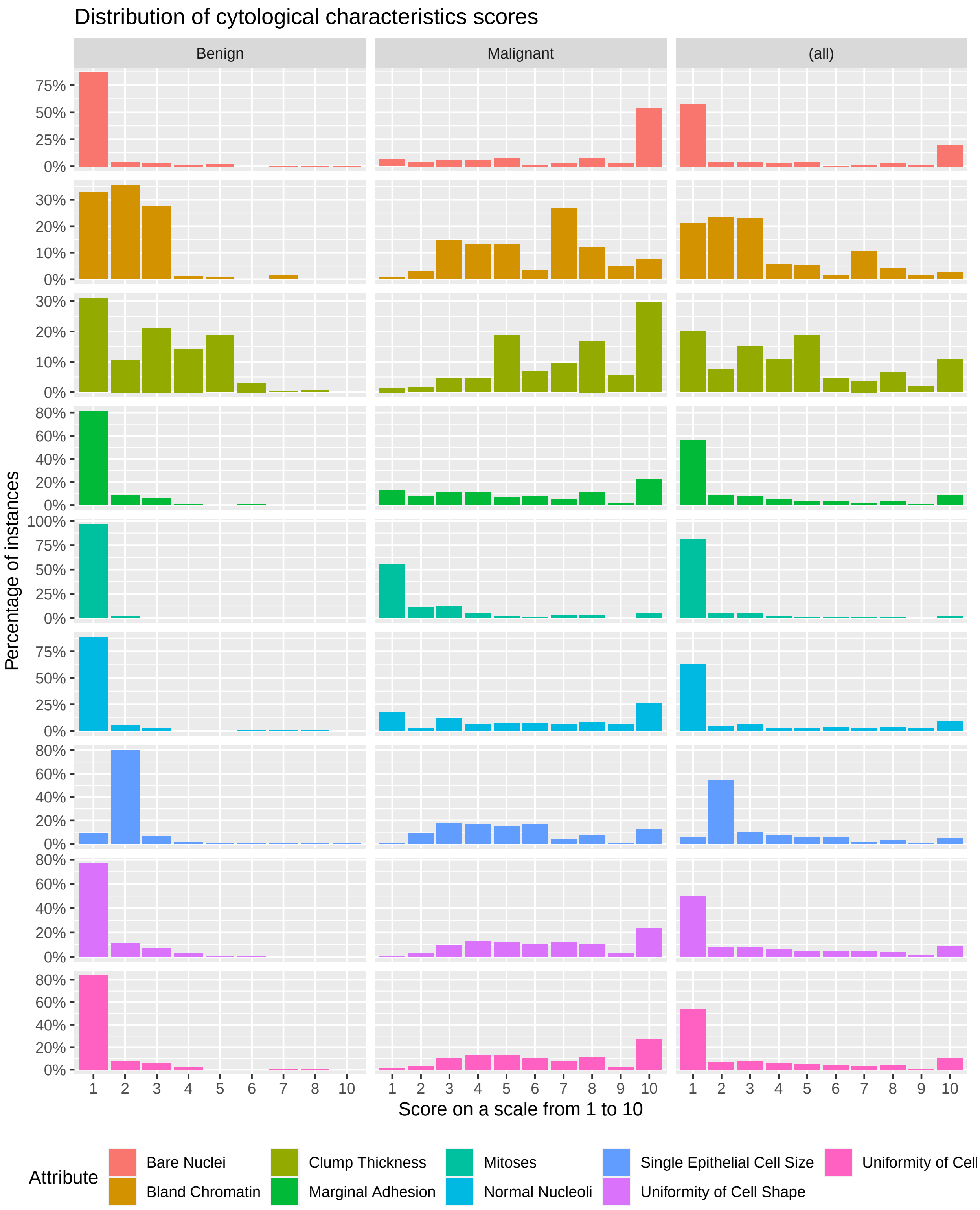

Figure 2: Distribution of data in percentage for 9 different cytological characteristics for benign instances, malignant instances and for all instances together

When looking at the distribution in *Figure 2* it seems there is a difference in distribution between the benign and malignant samples for every attribute. When looking at the overall distributions for the attributes it is important to keep in mind that the benign samples have a bigger influence on this because there are more instances with that class. The seemingly big difference in distributions for all of these attributes is a positive sign for machine learning, all of these attributes could be useful in differentiating benign and malignant samples from one another.

Now I will test if the differences observed in *Figure 2* are significant. I will use the Mann–Whitney U test for this as this is a non-parametric test that can be used for data that do not follow a normal distribution and ordinal data. For each attribute I will test if the malignant samples are significantly greater than the benign samples.

```
temp <- data.frame("attribute" = character(), "estimate" = double(), "pval" = double())

temp.data <- data

temp.data$class <- relevel(temp.data$class, ref = "Malignant")

for(attr in colnames(temp.data)[1:9]) {
  res <- wilcox.test(
    as.numeric(temp.data[,attr]) ~ temp.data$class,
    conf.int = TRUE,
    alternative = "greater"
  )

  full.name <- attribute.info$full.name[attribute.info$name == attr]
  temp <- rbind(temp, data.frame("attribute" = full.name,
                                 "estimate" = res$estimate,
                                 "pval" = res$p.value))
}

row.names(temp) <- NULL;

kbl(
  temp,
  col.names = c("Attribute name", "Estimate median of difference", "P value"),
  booktabs = T,
  digits = c(100),
  linesep = "",
  caption = "Results of one-sided Mann-Whitney U test for each attribute where the null hypothesis is th
) %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))
```

Table 49: Results of one-sided Mann–Whitney U test for each attribute where the null hypothesis is that the distribution of the malignant samples **is not** higher than that of the benign samples. And the alterernative hypothesis is that the distribution of the Malignant samples **is** higher than that of the benign samples. All of the p-values are well below 0.05 so we reject the null hypothesis for each attribute.

Attribute name	Estimate median of difference	P value
Clump Thickness	4.0000287732	1.736520e-68
Uniformity of Cell Size	5.0000208895	0.000000e+00
Uniformity of Cell Shape	5.0000340393	3.000000e-100
Marginal Adhesion	4.0000478861	1.197571e-77
Single Epithelial Cell Size	2.9999504915	1.339529e-84
Bare Nuclei	7.9999934047	8.290000e-98
Bland Chromatin	4.0000370221	3.375664e-79
Normal Nucleoli	4.9999897428	2.285017e-79
Mitoses	0.0000555406	1.003309e-39

```r
remove(temp, temp.data)
```

In *Table 49* the results of one-sided Mann-Whitney U tests are displayed. Each row of the table represents a Mann-Whitney U tests on an attribute. The null hypothesis for each test is that the distribution of malignant samples **is not** shifted to the right compared to the distribution of the benign samples. The alternative hypothesis is that the distribution of malignant samples **is** shifted to the right compared to the distribution of the benign samples. If the p-value for one such test is lower than 0.05 the null hypothesis is rejected. Since all the p-values are well below 0.05 the null hypothesis is rejected in all the tests. The estimate median of difference is the median of the difference between all the instances in the two samples. The size of the estimate median of difference of the bare nuclei is the largest. This is logical when looking at *Figure 2* you can see that most of the benign samples have a score of 1 and most of the malignant samples have a score of 10. The estimate median of difference of the mitosis also stands out. This estimate is very small, close to zero even. This also makes sense when looking at the *Figure 2* as you can see that for both benign and malignant cases most instances have a score of 1, but in the malignant cases the data is spread out more. So even though the change is not big there is a significant difference in distribution of mitosis between benign and malignant samples.

Relation between attributes

Now I will look at the correlation between the different attributes. My expectation is that some or all of the attributes will be positively correlated because the scores of the attributes range from 1 to 10, with 1 being the closest to benign and 10 the most anaplastic. This means that if the sample is benign, all or most of the scores will be 1 or close to 1, and if the sample is malignant the scores will tend to be higher than 1. This idea is also supported by the fact that for all the attributes the distributions of the malignant samples were significantly shifted to the right compared to the benign distributions, they all shift in the same direction. For calculating the correlation score I will not use the Pearson method, as it makes the assumption that the data is normally distributed, in this case the data is not normally distributed, so I will have to find a different method. I have chosen to use Kendall τ_b rank correlation as seems to be better at handling ties.

```r
df <- data

for(i in 1:9) {
  df[,i] <- as.numeric(df[,i])
}
colnames(df) <- attribute.info$full.name[-1]
```

```r
# Calculate  p values for correlation coefficients
correlation.p.values.kendall <- cor_pmat(df[,1:9], method = "kendall")

# Plot correlation coefficients for attributes
ggcorrplot(
  cor(df[,1:9], method = "kendall"),
  type = "lower",
  outline.col = "white",
  lab = TRUE,
  p.mat = correlation.p.values.kendall
) +
  ggtitle("Correlation between the attributes")
```

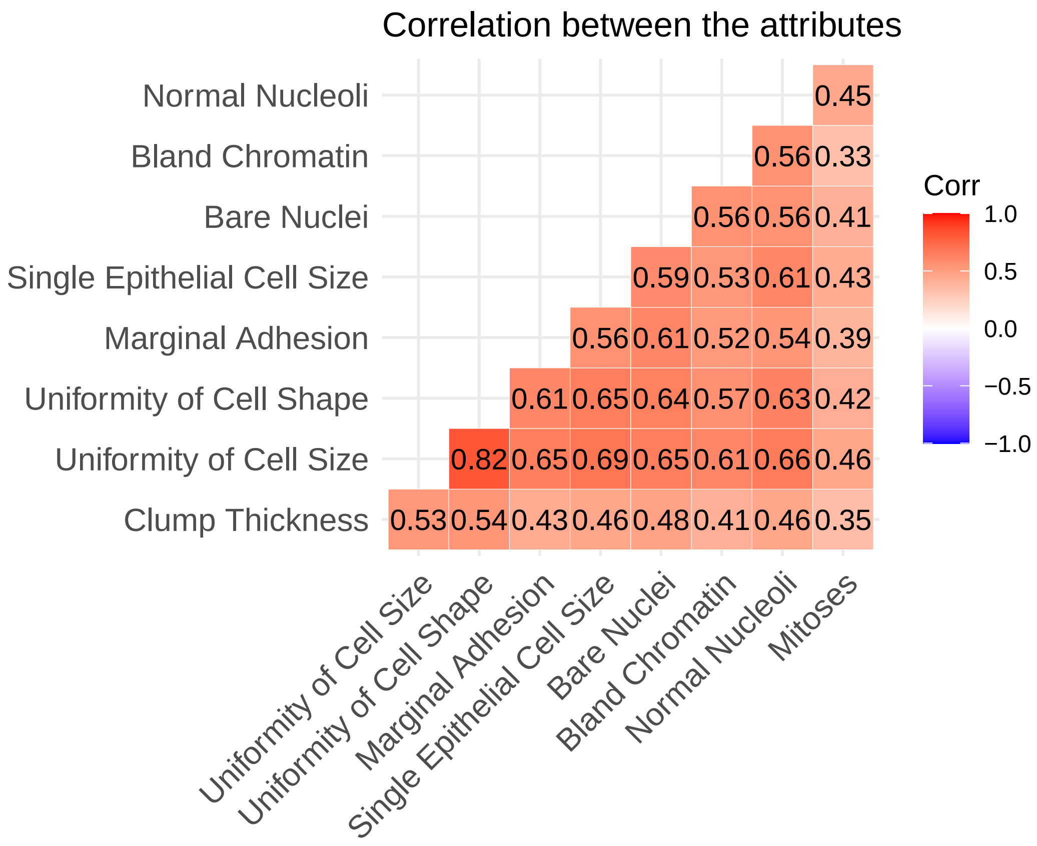

Figure 3: Correlations between the attributes

```r
# Print p values in table
kbl(
  correlation.p.values.kendall,
  booktabs = T,
  linesep = "",
  caption = "P values of Kendall $\\tau_{b}$ correlation coefficient",
  digits = 50,
  escape = F
) %>%
  column_spec(1:10, width = "3cm") %>%
  column_spec(1:10, width = "2cm") %>%
  kable_styling(latex_options = c("HOLD_position", "striped", "scale_down"))
```

Table 50: P values of Kendall τ_b correlation coefficient

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
Clump Thickness	0.000000e+00	0.000000e+00	0.000000e+00	9.225525e-43	6.300000e-49	2.000000e-50	1.051763e-40	3.149500e-46	8.093515e-26
Uniformity of Cell Size	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.999307e-39
Uniformity of Cell Shape	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.799174e-34
Marginal Adhesion	9.225525e-43	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	9.335989e-29
Single Epithelial Cell Size	6.300000e-49	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7.454364e-35
Bare Nuclei	2.000000e-50	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.748064e-31
Bland Chromatin	1.051763e-40	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.602932e-22
Normal Nucleoli	3.149500e-46	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.340816e-37
Mitoses	8.093515e-26	4.999307e-39	6.799174e-34	9.335989e-29	7.454364e-35	6.748064e-31	3.602932e-22	6.340816e-37	0.000000e+00

When looking at the correlations between the variables in *Figure 3* it is clear that most of the attributes have a moderate to strong relationship. Clump thickness and mitosis seem to have a weak to moderate relationship to the other attributes. The attributes uniformity of cell size and uniformity of cell shape stand out as having a very strong relationship with each other (with a Kendall's τ_b correlation coefficient of 0.82). As expected all the correlations are positive. When looking at the p-values corresponding to the correlation coefficients in *Table 50* it can be seen that the p-values are all very small meaning that all the correlation coefficients in *Figure 3* are statistically significant. These correlation coefficients could impact specific classifier learning algorithms. One such example is the Naive Bayes classifier, as it assumes that the values of the attributes are independent of each other.

Principal component analysis

Next I will conduct principal component analysis, so we can see if two distinct groups can be seen based on the principal components. I have some doubts whether or not I should scale the data before performing the PCA. All of the attributes are on a scale from 1 to 10. Before making the decision I will check the variance of the different attributes.

```
vars <- apply(data[c(-ncol(data))], 2, var)
attr.names <- attribute.info[attribute.info$name %in% names(vars),][,c("full.name", "name")]

var.per.attr.df = data.frame(attr.name = attr.names$full.name, vars = vars[attr.names$name])

kbl(
  var.per.attr.df,
  caption = "Variance per attribute",
  row.names = F,
  col.names = c("Attribute", "Variance"),
  booktabs = T,
  linesep = "",
  digits = 3) %>%
  kable_styling(latex_options = c("striped", "HOLD_position")) %>%
  column_spec(1:2, width = "7cm")
```

Table 51: Variance per attribute

Attribute	Variance
Clump Thickness	8.193
Uniformity of Cell Size	9.441
Uniformity of Cell Shape	9.017
Marginal Adhesion	8.569
Single Epithelial Cell Size	5.089
Bare Nuclei	13.463
Bland Chromatin	6.101
Normal Nucleoli	9.733
Mitoses	3.105

The variance of the attributes is displayed in *Table 51*. The variance is not the same for the attributes. This means that when performing PCA that the attribute with the highest variance will have the most influence on the principal components, and the attributes with the lowest variance will only have a little influence on the principal components. The question remains whether the attributes values are on the same scale. All the attributes are on an ordinal scale from 1 to 10, in this manual you can find a more extensive explanation how these are scored. When reading the table on page 4 of this document it becomes clear that the attributes all are scored on a scale from 100% normal to 100% anaplastic, but the steps between the percentages differ slightly (most steps of all the attributes are 10% difference, but some include steps of 15%, and some have 1 step of 20%). It is also hard to say if 10% difference in mitosis means the same as 10% difference in clump thickness for example. Another problem for a correct PCA is that not all the steps are of the size, in an ideal PCA every step would have the same meaning.

Now I will do the analysis. I have chosen to do the analysis twice, once with scaling and once without scaling and compare the results with each other.

```
scaled_options = c(TRUE, FALSE)

pca.list = list()
plot.list = list()

for(scaled_option in  scaled_options) {
  # Get principal components
  pca.res <- prcomp(df[1:9], scale. = scaled_option, center = TRUE)
  pca.list[[paste("scaled.", scaled_option, sep = "")]] <- pca.res

  # Calculate explained variance
  var.explained.df <- data.frame(
    PC= paste0("PC",1:9),
    var.explained=(pca.res$sdev)^2/sum((pca.res$sdev)^2)
  )

  # Plot explained variance for PC's
  new.scree.plot <- ggplot(var.explained.df, aes(x=PC,y=var.explained, group=1))+
    geom_point(size=4)+
    geom_line()+
    ylab("Variance explained") +
    xlab("Principal component")

  # Get points to plot in PCA plot
  df.pca <- data.frame(pca.res$x, class=data$class)
  df.benign <- df.pca[df.pca$class == "Benign", ]
  df.malignant <- df.pca[df.pca$class == "Malignant", ]

  # PCA plot
  new.pca.plot <- ggplot(df.pca, aes(PC1, PC2, col=class)) +
    geom_point() +
    coord_cartesian(xlim = 1.2 * c(min(df.pca$PC1), max(df.pca$PC1)),
                    ylim = 1.2 * c(min(df.pca$PC2), max(df.pca$PC2))) +
    geom_encircle(data = df.benign) +
    geom_encircle(data = df.malignant) +
    xlab("Principal component 1") +
    ylab("Principal component 2") +
    theme(legend.direction = "horizontal", legend.background = element_rect(linetype = "solid", size =

  leg <- get_legend(new.pca.plot)
  new.pca.plot <- new.pca.plot + theme(legend.position = "none")

  plot.list[[paste("scree.plot.scaled.", scaled_option, sep = "")]] <- new.scree.plot

  plot.list[[paste("pca.plot.scaled.", scaled_option, sep = "")]] <- new.pca.plot

}

# Add legend to plotlist
plot.list[["leg"]] <- leg

# Titles for rows and columns wrapped plot
row1 <- ggplot() +
  annotate(
    geom = ’text’,
    x=1, y=1,
    label="Scaled = TRUE",
    angle = 90,
    size = 5,
    fontface = 2) +
  theme_void()
row2 <- ggplot() +
  annotate(
    geom = ’text’,
    x=1, y=1,
    label="Scaled = FALSE",
    angle = 90,
    size = 5,
    fontface = 2) +
  theme_void()
col1 <- ggplot() +
  annotate(
    geom = ’text’,
    x=1, y=1,
    label="Explained variance",
    size = 5.5,
    fontface = 2) +
  theme_void()
col2 <- ggplot() +
  annotate(
    geom = ’text’,
    x=1, y=1,
    label="PCA plot",
    size = 5.5,
    fontface = 2) +
  theme_void()

title.list <- list(a = row1, b = row2, e = col1, f = col2)

layoutplot <- "
#ccccdddd
aeeeeffff
aeeeeffff
aeeeeffff
bgggghhhh
bgggghhhh
bgggghhhh
#####oooo
"

wrap_plots(plotlist = c(title.list, plot.list), guides = ’collect’, design = layoutplot)
```

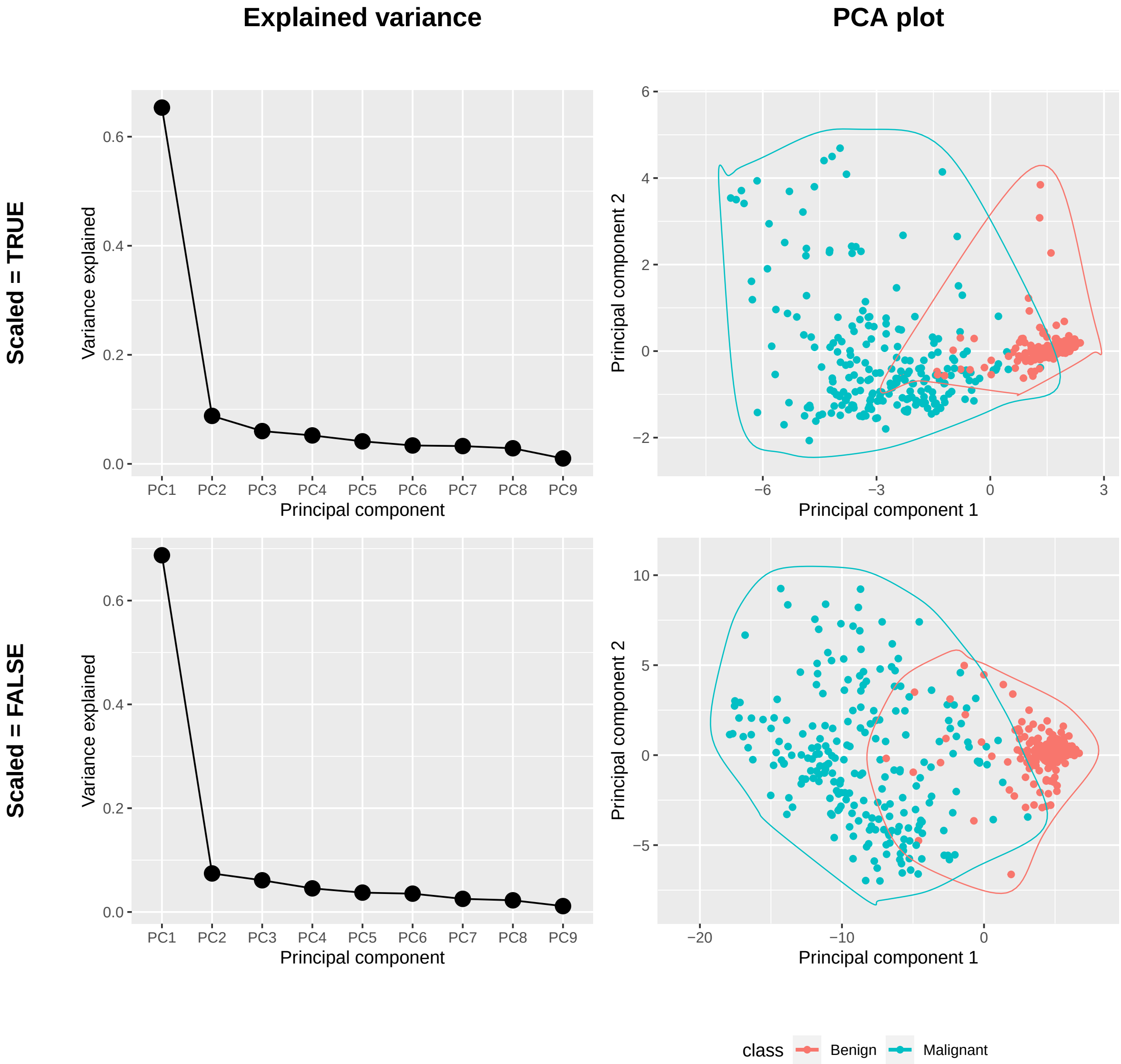

Figure 4: Principal component analysis

```r
kbl(
    pca.list[["scaled.TRUE"]]$rotation,
    caption = "PCA: Rotation matrix of the 9 cytological characteristics to each principal component, s
    booktabs = T,
    linesep = ""
  ) %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position")) %>%
  column_spec(1:9, width = "2cm")
```

Table 52: PCA: Rotation matrix of the 9 cytological characteristics to each principal component, scaled = TRUE

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Clump Thickness	-0.3026067	-0.1277601	0.8730402	-0.0509633	0.0200015	0.2080380	0.0594103	0.2830144	-0.0018746
Uniformity of Cell Size	-0.3821300	-0.0320382	-0.0386545	0.1856231	-0.1336961	0.2422388	-0.1310226	-0.4164032	-0.7415439
Uniformity of Cell Shape	-0.3780364	-0.0701172	0.0287437	0.1593851	-0.0888896	0.1701730	-0.0535618	-0.5990616	0.6537113
Marginal Adhesion	-0.3333947	-0.0734380	-0.3931759	-0.5003666	0.0309380	0.5636711	0.3050514	0.2529206	0.0528526
Single Epithelial Cell Size	-0.3360831	0.1847634	-0.1429394	0.3430707	-0.7025839	-0.1980931	0.1571675	0.3895393	0.0739862
Bare Nuclei	-0.3335249	-0.2678899	0.0273743	-0.5168253	-0.0492067	-0.6806530	0.1963920	-0.1930489	-0.0871259
Bland Chromatin	-0.3465219	-0.2393446	-0.1911542	0.0159296	0.2005512	-0.1029457	-0.7827851	0.3403353	0.0802731
Normal Nucleoli	-0.3355534	0.0240361	-0.1277002	0.4830818	0.6411883	-0.1749484	0.4228656	0.1273752	-0.0195346
Mitoses	-0.2268878	0.8992081	0.0828547	-0.2621999	0.1596186	-0.0873072	-0.1689347	-0.0510517	0.0093337

```r
kbl(
    pca.list[["scaled.FALSE"]]$rotation,
    caption = "PCA: Rotation matrix of the 9 cytological characteristics to each principal component, s
    booktabs = T,
    linesep = ""
  ) %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position")) %>%
  column_spec(1:9, width = "2cm")
```

Table 53: PCA: Rotation matrix of the 9 cytological characteristics to each principal component, scaled = FALSE

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Clump Thickness	-0.2984508	-0.0683715	0.8550914	-0.1235371	0.3674030	-0.1086196	-0.0643288	-0.0941102	0.0056192
Uniformity of Cell Size	-0.4011692	0.2316964	0.0016082	-0.2542141	-0.3547529	-0.1012452	0.1872058	0.1000193	0.7345765
Uniformity of Cell Shape	-0.3903768	0.1644123	0.0625476	-0.1638881	-0.3824158	-0.0908336	0.4185489	0.1888991	-0.6514939
Marginal Adhesion	-0.3361819	-0.1293465	-0.4820717	-0.4363429	0.6011305	-0.2200799	0.1288686	-0.1373377	-0.0469165
Single Epithelial Cell Size	-0.2501787	0.1993757	-0.0558834	-0.1757126	-0.2191651	0.4036996	-0.3528604	-0.7162597	-0.1229510
Bare Nuclei	-0.4388940	-0.7711234	-0.0686198	0.3265365	-0.1363697	0.2707260	0.0675639	0.0130594	0.0691472
Bland Chromatin	-0.2920065	0.0021110	-0.1107320	0.0551576	-0.1433030	-0.4273832	-0.7704391	0.2996685	-0.1133753
Normal Nucleoli	-0.3590016	0.4785523	-0.1097836	0.7218032	0.2996154	0.0001368	0.1198952	-0.0650426	0.0260471
Mitoses	-0.1233318	0.1856918	-0.0191080	-0.2032637	0.2337317	0.7087615	-0.1788822	0.5651236	-0.0170201

Now it is time for the fun part: interpreting the results of the PCA's. In the PCA plots in *Figure 4* it can be seen that with or without scaling, the benign and malignant samples clearly form two distinct clusters, this is a positive sign for making a classifier for this data. The in the unscaled plot the clusters might seem a bit more distinct than in the scaled plot.

In the scaled PCA plot (top right) the variance in the second principal component seems to represent a small amount of instances. This second component does seem to say something about malignancy, you can see that only 3 benign instances are higher than -2, a lot more malignant instances are in this area (even though benign is the majority class). Looking at *Table 52* we can observe that the second component is dominated by the mitosis attribute. Comparing the second principle component to the distribution of mitosis in *Figure 2* the distribution of the second component makes sense, most of the instances have a mitosis of 1, but from the instances that have a mitosis score higher than 1 most of them are malignant.

In the unscaled PCA plot (bottom right) principal component 1 seems to have a dividing line between benign and malignant at around -2.5. On principal component 2 the benign instances seem to be concentrated around 0 and the malignant instances are more dispersed.

Comparing the explained variance plots in *Figure 4* you can see that they are quite the same. In both cases the first component explains by far most of the variance. In the unscaled PCA the first component explains a little more of the variance than the scaled version and the second component a little less. The difference in these plots doesn't seem significant. It is important to keep in mind that in the unscaled version the different attribute do not contribute equally to the total variance.

When comparing *Table 52* and *Table 53* you can see that in *Table 52*, the scaled PCA, the attributes contribute quite equally to principal component 1, the smallest contributor to component 1 is mitosis, but mitosis does dominate the second component. This is explained by mitosis having a smaller relationship to all the other attributes, as can be seen in the correlation plot (the greater the correlation between two attributes, the more they can be displayed on the same axis). In *Table 53*, the unscaled PCA, the attribute with the biggest variance (bare nuclei) contributes most to the first as well as the second principal component and mitosis does not play a big role in either of the two.

Keeping the correlations between the attributes in mind, and the fact that all the attributes contribute more equally I think that the PCA plot with scaling is the better visualization of the multidimensional data, but the unscaled plot seems to have a better division between benign and malignant instances, this could be caused by the fact that the attribute with the biggest variance is also the best predictor of malignancy. However this is not the core purpose of PCA. If we want to find the axis that represents the biggest separation in class we should do *linear discriminant analysis.*

Machine Learning Phase